New England Journal of Public Policy

Volume 17 | Issue 2 Article 7

3-21-2002

School District Performance Under the MCAS

Jie Chen
University of Massachusetts Boston, jie.chen@umb.edu

Thomas Ferguson
University of Massachusetts Boston

School District Performance Under the MCAS

Jie Chen
Thomas Ferguson

Education reform has spawned efforts to test learning across the nation. This paper analyzes the determinants of Massachusetts' school district test scores under the state's high stakes testing program, MCAS. The study is the first to demonstrate direct links between improvements in MCAS scores and state aid to school districts. The authors estimate "value added" for each school district in the state. The list of schools with high value added produces real surprises — while some affluent districts do well, others rank at the very bottom. Additionally, the study analyzes how teacher maximum salaries, district superintendent salaries, per capita income, internet usage, and other factors, including various state school "choice" initiatives, affect test scores. The study shows that athletic budgets have substantial impacts on district test scores — the more districts spend on athletics as a percentage of the state's "foundation budget," the lower their scores.

This is the first study to cover the entire state and employ appropriate spatial statistics to correct well-known errors in statistical estimates of geographic data.

> From the battlefield adjutants were continually galloping up to Napoleon with reports from his marshals of the progress of the action. But all those reports were deceptive; both because in the heat of battle it is impossible to say what is happening at any given moment, and because many of the adjutants never reached the actual battlefield, but simply repeated what they heard from others, and also because, while the adjutant was galloping the two or three versts to Napoleon, circumstances had changed, and the news he brought had already become untrue….
>
> Upon such inevitably misleading reports Napoleon based his instructions, which had mostly been carried out before he made them, or else were never, and could never, be carried out at all….
>
> For the most part what happened was the opposite of what they commanded to be done. The soldiers ordered to advance found themselves under grapeshot fire, and ran back. The soldiers commanded to stand still in one place seeing the Russians appear suddenly before them, either ran away or rushed upon them; and the cavalry unbidden galloped in after the flying Russians….
>
> In reality all these movements forward and back again hardly improved or affected the position of the troops.[1]
>
> —Tolstoy, *War and Peace*, Chapter 33

Jie Chen is Statistician, Computing Services, at University of Massachusetts Boston. Thomas Ferguson, professor of political science at University of Massachusetts Boston, is a member of the National Advisory Panel of the Quality Initiative of the Council for Aid to Education.

In the shadow of 9/11, once fashionable allusions to America's "great school wars" now sound unbearably glib. But as one contemplates the current political struggles over the state's high stakes testing program — the famous MCAS (Massachusetts Comprehensive Assessment System) — it is frequently hard not to recall Tolstoy's celebrated account of the Battle of Borodino. While average scores in most districts are rising, a radical change in the scoring of the English test for 2001 clouds the real extent of improvement. Nor, almost a decade after the landmark Massachusetts Education Reform Act, has anyone succeeded in resolving the Commonwealth's modern Riddle of the Sphinx: whether there is any relation between the test score changes and the state's prodigious investments in K-12 education.[2] So far, the boldest claim is that increased state spending may have elevated fourth grade scores (only) in the final years of MCAS's low stakes predecessor, the Massachusetts Educational Assessment Program (MEAP). Other researchers who have pored over data for MCAS itself report mixed or negative findings and they counsel patience, until the passage of time secures more data.

Analyzing Change

We will see below that the analysis of "change" is among the most treacherous areas in all of statistics, one that has in recent years claimed the attentions of many gifted researchers. Thus, even amid tightening state budgets, it is relatively easy to understand why advice simply to let the taxi meter run has yet to inspire overt resistance. But more than the big picture is obscure. Just as at Borodino, facts in the field are murkier than almost anyone realizes: While test results for individual pupils are not, of course, released publicly, the Department of Education (DOE) publishes scores aggregated by school and district, which anyone can download from its website with simple equipment. Yet early in our research, we were taken aback to discover that, using this data, we could not replicate the scores reported in the major media, or the rankings of school districts derived from them.

Several reviews of MCAS data have already been published, or at least discussed in newspapers.[3] Each makes useful points, but all are brief and consider only a handful of issues. The Department of Education's "Technical Reports," available on its website, provide more extensive analysis, but are also narrowly focused. If the DOE or the test maker has conducted longer studies, they have not been made public.[4] In the meantime, impressions persist that "demography is destiny" — that school characteristics have little to do with test outcomes and that children in affluent school districts can hardly help scoring well, simply because they come from affluent backgrounds. Data indicating drastic ethnic and racial disparities in scores also fan anxieties about possible discrimination. Many also claim that competition from new, state-supported charter schools must inevitably spur public schools to higher standards of performance and perhaps even mitigate existing racial and ethnic disparities. And, of course, a large segment of popular opinion wonders if the tests reliably measure anything at all, even as many homebuyers and not a few school boards, superintendents, citizens, and real estate agents track year-to-year fluctuations in test scores as though they were major league batting averages.[5]

Some of these views are rooted in demonstrable misunderstandings and do not require extensive analysis. For example, opponents of high stakes testing perform a valuable service when they remind everyone that individual test scores are often highly inaccurate. But it simply does not follow that test results for large groups of

students are nonsense. The decisive point is that some scores come in too high and others, which tend, not surprisingly, to garner more publicity, come in too low. In large groups, such cases normally wash out, making group averages far more accurate than results for individuals.

A Study in Four Parts

But making real progress on other MCAS data fronts is not so easy. It requires a lengthy review of the evidence, with more sophisticated statistical techniques than have thus far been applied. This paper presents a detailed quantitative assessment of many issues connected with the MCAS. In Part I, we begin by clarifying questions about the data. We show that DOE data as used by major newspapers in the Commonwealth to identify best performing districts fail to do so, for uncontroversial statistical reasons. We correct the error and produce (in Appendices 1 and 2) a list of the best performing districts based on the same DOE data.

Part II of our paper analyzes the determinants of school district performance under the MCAS. Changes in scoring initiated in 2001 do not prevent comparisons with earlier years, but the switch does mean that without access to additional data, the range of statistical techniques that can be employed is relatively meager. Our aim in this paper is to dissect the state's experience with MCAS in as much detail as possible, so that we can shed light on the most urgent public policy questions the test raises. Accordingly, we consider results for the test from the years 1998, 1999, and 2000, when scores are generally agreed to be directly comparable. Our consideration of results for 2001 is brief, limited to Appendix 2's district grand averages for reference purposes.

We begin by developing a "cross-sectional" model of district success. In this traditional approach time plays no essential role. The aim is to understand the determinants of each school district's consolidated grand average score over those three years as a whole. Sorting through a wide range of specifications and potential influences, we try to pick our way through the statistical problem of multicollinearity, or the fact that some variables that affect student performance are so closely associated that they are all but impossible to distinguish. Eventually we arrive at a plausible, statistically rigorous model of district performance.

We employ this model to assess a much broader range of potential influences on school district tests than previous studies, including the salaries not only of teachers, but of school superintendents, the influence of sports, curriculum (in certain limited measures), district political competition, the availability of computers, and the allocation of money across various budgetary categories, such as spending on teachers' aides. We also inquire whether the many claims made about the salutary effects of school "choice" and charter schools square with the facts of performance on the test and whether district participation in the METCO program that buses children from inner cities to suburban schools affects scores.

Part III of our paper attempts to demonstrate that MCAS can be employed to assess "value added" and not simply to ratify claims of the state's most affluent communities to harbor the "best" schools in the state. We employ our model to identify schools that consistently score above what their economics and demography predict. The resulting roster of "best performing" districts is strikingly different from conventional lists based simply on raw scores. It yields real surprises and cannot be dismissed as a simple reflection of social circumstances. It also is proof against objections advanced by some members of the Massachusetts Department of Education

that test scores from these early years cannot be used to rank districts because students who were not required to pass the MCAS to obtain a diploma may not have taken the test seriously enough.[6]

At last in Part IV, in almost Kirkegaardian fear and trembling, we tackle the question of change — of what might be styled the "billion dollar question" of whether the state really has anything substantial to show for all the money it has lavished on K-12 education. To answer this query, we employ a relatively new statistical technique, the so-called "multi-level" or "growth curve" approach to analyzing longitudinal data with repeated measures on the same subjects (here, school districts). Our conclusions are necessarily shadowed by many qualifications that are inherent in the analysis of educational outcomes over time. They can also easily be overstated, since the effects in question (which are almost certainly still evolving) appear thus far to be modest.

Methodology

Still, in the end, our results are fairly clear cut: While we detect signs of real inefficiencies in the system, we also find solid evidence that the increases in school district MCAS scores are positively related to increased state funding for K-12 education. Our results also contain surprising implications regarding racial and economic disparities in the MCAS. We find, for example, that while the percentage of African-Americans in a district influenced the district's starting position in the MCAS process, it appears to have no effect on subsequent rates of improvement. District rates of improvement also appear to be independent of the presence of other minority groups and of current per capita income. But using recently released data from the 2000 U.S. Census, we do find that rates of improvement on the test are related to changes in district household income between 1989 and 1999.

Our paper contrasts with previous studies in other ways besides the range of the questions we address. For example, we are clear that MCAS data are spatial data. That is, test scores are not simply numbers drawn from random samples, but come from particular school districts located in definite places. Geographers have long recognized that such data are highly likely to exhibit "spatial autocorrelation" — a pattern in which areas physically adjacent to one another strongly resemble each other. Such data cannot reliably be investigated with the customary statistical techniques of "ordinary least squares" regression. They require special methods that are all but unknown in educational statistics and not all that common even in econometrics.[7] In this paper, we rely heavily on Geographic Information System (GIS) techniques used to map data that is displayed spatially.[8] These techniques calculate a neighborhood matrix that specifies distances between districts to test for the presence of spatial autocorrelation. Where this is present — as it is in virtually all our cross-sectional results — we correct for it. But while the methods used in this paper are admittedly complex, we relegate purely technical details to Appendices or notes. Our main text is designed to be read by anyone concerned with the policy issues at stake in the MCAS debate.

We try to improve on previous studies in a second, major way by recovering data that they drop out. In Massachusetts most school districts represent towns. But a fair number of smaller, sparsely populated municipalities share part of their school systems — normally the (relatively expensive) high school, and frequently part or all of the middle school — with one or more neighbors. The shared part of the system

normally constitutes a separate school "district" for reporting purposes. Thus towns like Acton and Boxborough each run a K-6 local system while jointly administering the overlapping Acton-Boxborough district from grade 7 on.

For statistical studies such arrangements create real headaches. The MCAS tests we study in this paper were administered to students in grades 4, 8, and 10.[9] Comparing grand averages of districts with and without high schools is like comparing apples with oranges. Consequently, previous studies have simply dropped the incomplete districts.[10] But throwing so much data out is dangerous. Also, the number of districts that are usually excluded for this reason is not negligible — depending on how one counts, they may amount to a quarter of the whole sample.[11] When the size of the sample is reduced, the reliability of results decreases, and the task of sorting out the operative variables — the problem of statistical multicollinearity — becomes all the more intractable. It is no way to investigate a process with such sweeping implications for so many people.

We restore the data lost to previous studies by combining results for the component districts. For example, in our dataset the lower grades in Acton and Boxborough are combined with the joint Acton-Boxborough upper grades to create one "Acton-Boxborough" district. Since the "units," once conglomerated become larger, many variables, particularly from the Census, require reweighting before they can be employed. This is both tricky and laborious; it is easy to see why previous analysts preferred simply to drop these cases. The result, however, is that our study is the only one that actually reports on data for the entire state.[12]

But while we are confident that our study sheds new light on MCAS, it is also necessary to point out some limitations. First, the data we analyze represent results from what are now widely known as "quasi-experiments." Like astronomers or anthropologists, we are limited to observing variables in combinations found "in the wild." In contrast to laboratory scientists, we cannot design experiments to help sort co-occurring phenomena. This point may sound like a truism. It is not. Its sweeping implications for statistical studies have only recently been widely taken to heart by methodologists. Since one cannot manipulate, one cannot randomize the objects one is studying. And since we have only a few years of data, some techniques one might use to get around this problem are unavailable. It is thus difficult to rule out variables that are really irrelevant but happen to be strongly correlated with what one is studying or, conversely, to recognize the significance of hidden variables that are subtly correlated with the more obvious variables one is studying.[13]

Further Cautions

Statistical multicollinearity almost always aggravates these problems; for example, income, race, ethnicity, poverty, and many other variables — such as school dropout rates — are frequently so highly correlated that they cannot easily be distinguished statistically, even in large samples. This makes for potentially mistaken analysis because it inflates the variance of one's estimates and thus muddies results. (The other horn of this dilemma also has potentially important policy implications: Separately reporting scores by race or gender, without controls for the omitted variables, as both the DOE and some academic studies have done, overstates the importance of the highlighted variable since it appears to be responsible for all the variation.) Techniques exist for recognizing and treating multicollinearity, and we have made heavy use of them, since we are extremely reluctant to rely on the most common

approach — which is simply to drop most of the offending variables — without some evidence about which works best or misleads the least.[14] Nevertheless, the reader is warned.

Other yellow caution flags need to be raised as well. Our data for income and related economic magnitudes treat the state as having a uniform price level as if it is no more or less expensive to live in one town or another. This is surely an oversimplification, but price deflators for individual towns or counties do not exist, so there is no practical alternative. Note, however, that formulas for state aid do claim to make allowances for differing prices, so some corrections are built into the data.[15]

It is also important to note that our data are for entire school districts. They thus aggregate results for groups of groups of groups: students nested in classes, within schools, inside districts. We work only on district data. Once again, the problem this situation generates — that of so-called "cross level" inference — has received much attention in recent statistical literature. Still the major pitfalls bear repeating: Data this highly aggregated display much less variability than do individual scores. So an apparently minor difference in scores between districts can signal a yawning gulf, rather than a minor difference, although many individual scores within the lower scoring unit will still tower over many in the higher.

But the most important warning concerns the "ecological fallacy." If units at one level of analysis, say, districts, display a correlation between scores and some characteristic (such as ethnicity), it does *not* necessarily follow that this correlation holds true at other levels; for example, between schools within districts or within particular grade levels or classrooms. Indeed, an exactly opposite pattern may hold at other levels. An example from political science illustrates how easily the best laid plans of a whole discipline can go astray: Before the advent of modern sampling techniques, political analysts frequently tried to descry trends in ethnic voting by focusing on precincts that were heavily dominated by particular voting blocs. Now everyone recognizes that voters living in such peculiar niches may be wildly unrepresentative of their ethnic group as a whole.[16]

The concentration on district-level outcomes affects our results in other, subtler ways. Some of the apparent variation among districts really represents variation between schools or neighborhoods within these districts as well as, quite possibly, between teachers. These factors, however, are not modeled in this paper. Instead, all variation appears as variation across districts. Capturing variations below this level — at the level of the schools or classes — would require substantial reconfiguration of the GIS data available to us. We have, accordingly, reluctantly put off the investigation for another time.[17]

Though we gathered new data on some key variables ourselves, such as superintendents' salaries, there are other data we would still like to have.[18] For example, recent evidence suggests that school grading practices influence student performance, and thus, presumably, MCAS results.[19] We also suspect that the widespread practice of starting high school classes very early in the morning to economize on transportation costs by employing school buses in shifts seriously affects student achievement. But the Department of Education does not publish data on these topics. Given our focus on districts, we have not found any way to investigate cases where there is reason to suspect that the MCAS exams were compromised in a particular school.[20] Like the troops on all sides at the battle of Borodino, accordingly, we try to make sense of the situation as we find it.

I. What We Study: The Dependent Variable and Data

At Borodino Tolstoy claimed that Napoleon could not tell "where what he had seen was" when he lifted his eyes from his spyglass and "looked again with the naked eye."[21] Something like this confusion exists with regard to MCAS scores. As mentioned, the Department of Education reports the results of the tests for both schools and districts. But the DOE does not consolidate results for the separate tests — English, Math, Science, and the like — into a single "grand average" for each grade level. Neither does it publicly aggregate results for all grade levels and schools into district average scores.

But these are easily calculated from the data. If one believes, as virtually everyone does (including ourselves), that every grade level is equally important and that results for separate tests should figure equally in the overall average, then the task of computing grand averages for districts or schools is akin to computing the average of a single test with nine parts.[22] (Nine, because each district in the period with which we are concerned administered three different tests to students in the fourth, eighth, and tenth grades.) It is only necessary to add up the scores on each part and divide the resulting total by the number of children who took each test. The result will be what everyone recognizes as the average (mean) score for the district or school.[23]

The DOE does not, of course, list individual scores. But it does post average scores for each particular test and the number of children who took it. Multiplying these average scores by the number of pupils taking each particular test and then summing the resulting figures yields the same figure one would arrive at if one worked from original, individual test scores. Since the same total results, the number can, once again, be divided by the total number of students taking each test to arrive at an overall average.

Here, however, lies a snare celebrated in the lore of statistics. Since the tests are given on different days, the number of students taking each varies, sometimes substantially.[24] If one worked from individual test results, this would become obvious during the count. It is much less obvious if the data list the number of students taking each test and the corresponding average test scores in separate columns, as the DOE data does. In this latter circumstance, it is insidiously easy to make the mistake of trying to calculate averages for each unit without reference to the varying numbers of students who take each test. One is strongly tempted simply to add the "unweighted" test scores for each grade level. The combined total for all grades can be reported by itself as a sort of summary number; or, if one wants to calculate what looks like the "average," one can, once again, succumb to temptation and divide the combined total by the number of tests. However one does it, the results will be wrong.

Every published school ranking we have seen, including those in the *Boston Globe*, relies on these "unweighted" averages.[25] In Appendix 1, accordingly, we rank each district correctly based on the same DOE data used by the *Boston Globe*. We display results for each year, along with a consolidated "grand average" for all three years.[26]

The new results are invariably interesting, even when they are not earthshaking. In 1998 and 1999, the *Boston Globe* hailed the schools of the town of Harvard as the best in the state.[27] Correctly calculated results suggest that the Harvard system is

indeed outstanding. But in both 1998 and 1999, the Wellesley school system narrowly topped it. By no means are all the corrections this minor, particularly when one traces year-to-year changes in relative standings. Some shifts are quite large. For example, if one compares our results for 1999 with those presented in the *Globe*, some districts fall as much as 16 places. While we have not bothered to extend the comparison, a similar error in the opposite direction in the following year would imply a comparatively huge reshuffling.

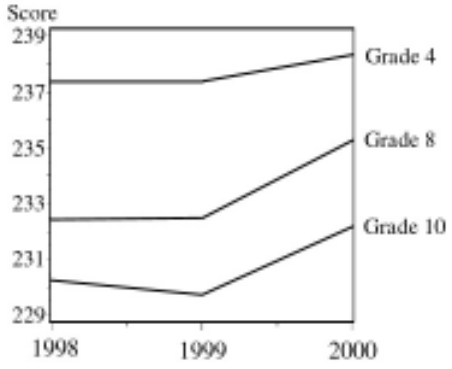
The last column of figures in our Appendix 1, the "Grand Average 1998–2000," supplies the variable we analyze in the next section of the paper. The figure for each school district reflects the combined weights of all three tests (math, science, and English) for the years 1998, 1999, and 2000. These averages reflect the work of so-called "regular" students — those who are not handicapped or otherwise specially disadvantaged.[28] The DOE separately reports scores for a wide variety of special student populations, including most English as a second language (ESL) students, as well as for vocational-technical school students. We believe that all of these merit careful attention — the question of who does well teaching handicapped students is, after all, a first-order problem of public policy, as is the question of how MCAS can be fairly administered to students in vocational schools, where curricula often differ by design from curricula in other schools. But test scores for such students cannot sensibly be analyzed using a model designed for the majority of students. In this paper, we are concerned only with these latter.

II. Modeling the Determinants of School District Success in the MCAS: Cross-Sectional Results

In the spirit of the French and Russian commanders who began their preparations for the battle of Borodino by surveying the terrain, we first look at some plots of the test score data. Figure 1 presents a year-to-year overview of the scores by grade level, before they are rolled into the grand averages that we analyze. It is immediately apparent that the fourth grade scores run somewhat higher than scores in the eighth or tenth grades. This figure, however, tells us only about statewide average

Figure 1

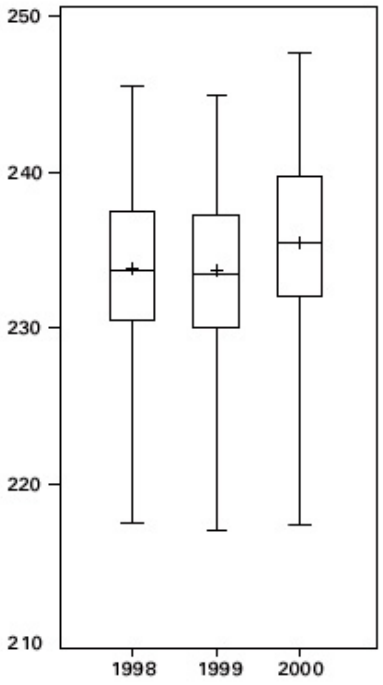
Test Score Trends



scores. It gives no clues to variations between school districts. Figure 2, which statisticians will at once recognize as a so-called "box plot," does just this. The length of the thin vertical lines for each year's scores marks the range of variation among district averages. They stretch from a low of just under 220 to a peak of a bit less than 250. The long rectangles between each pair of lines enclose, by convention, the middle half of all scores. Comparing the midpoints of each rectangle, we see that in 1999 the average score declined very slightly. Scores then rose rather substantially in 2000.

Figure 2

District Scores
Variation in Range and Average

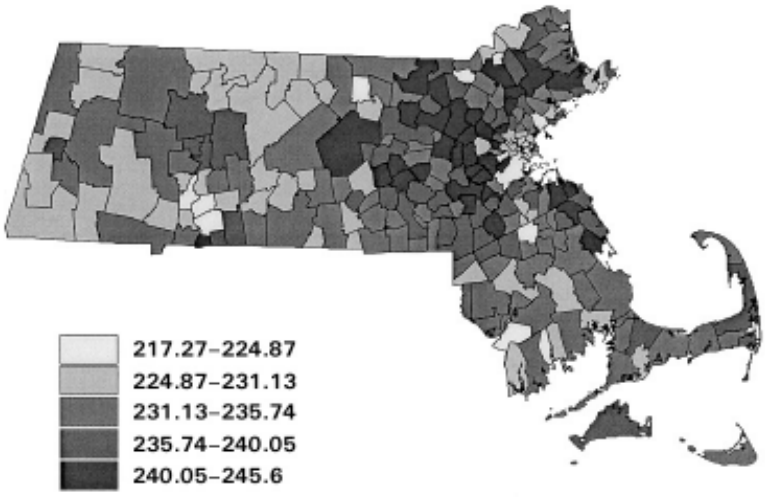


The geography of the test scores is interesting and important. Figure 3 plots all district grand averages on a state map, using GIS technology. This map is fairly brimming with implications for individual communities. The low scores of the state's large cities are apparent, as is the belt of high scoring suburbs encircling Boston along Route 128. The moderately low scores of rural districts are also easy to spot. So is another fact that is crucial for our paper: Neighborhood effects appear to be powerful. That is, many districts strongly resemble their neighbors. (See the Moran Tests for spatial autocorrelation, presented in Appendix 3, which confirm the judgment of the eye.)

The patchy geography of MCAS scores throws a shadow over the technique most commonly employed in educational statistics — that of ordinary least squares

Figure 3

Grand Average District Scores



regression. For, alas, least squares regression requires inflexibly that the residuals — the errors in values predicted by our model — not be correlated with each other. That is, in plain English, the predictive error for Wellesley needs to be random with respect to the errors for neighboring towns such as Weston, Sherborn, Newton, or Dover. The map warns that it obviously will not be, since their scores all fall within a narrow range.

Previous studies of MCAS have ignored this point. Indeed, none appear to have run any test for autocorrelation at all. But since MCAS scores are highly correlated across districts, there is no real alternative to adopting spatial regression techniques. These techniques require special software, but are otherwise not onerous. Once the autocorrelation is removed, one can proceed pretty much along customary lines, though the need to calculate spatial weights for every district complicates some econometric procedures and renders impossible many common tests of the effects of outlying data points. The difference this makes in the residuals for the resulting equation is easy to see if one compares the residuals of our spatial regression model with the non-spatial version in our Appendix 3.

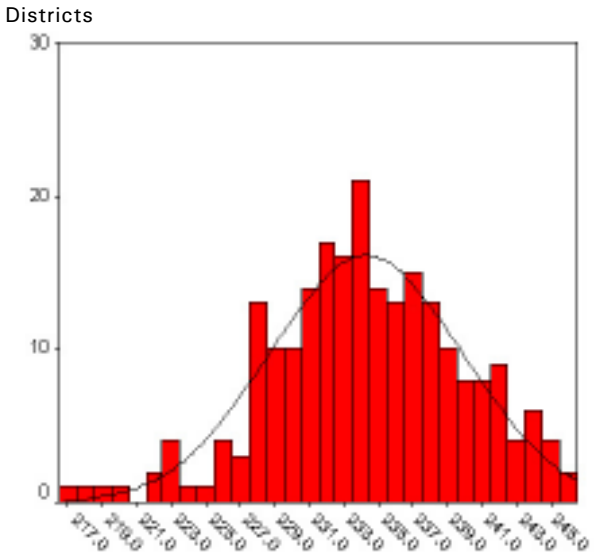
A plot of the values of our dependent variable — the three year district grand average — reveals something else. In a world in which the "bell curve" has the power of a symbol on a par with the cross or the American eagle, the MCAS — like most other well designed tests — rings in tones all its own. Figure 4 shows the distribution of scores for the three year grand average. While the unweighted scores could plausibly be claimed to resemble a normal curve (1), this claim is undercut if the scores are appropriately weighted (2).[29]

A close look at the right side of Figure 4(2) shows almost no high scoring districts. Since the Department of Education has now changed the scoring system for the English test, we can simply state the obvious: scores on the fourth grade English

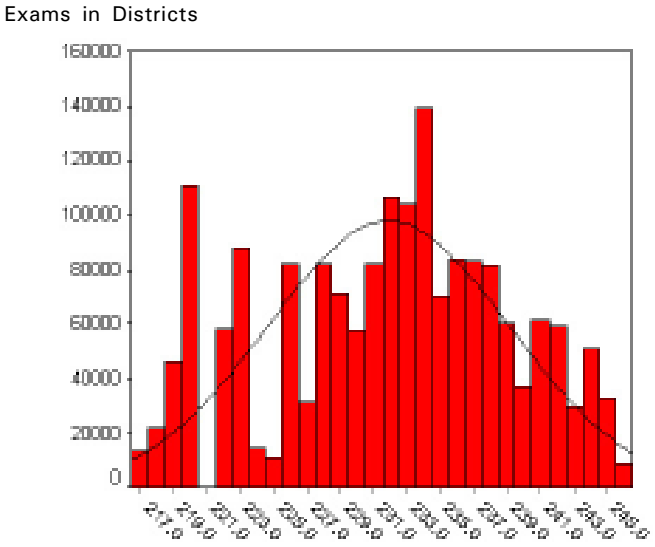
Figure 4

Distribution of Scores, Grand Average for 1998–2000

1. Unweighted



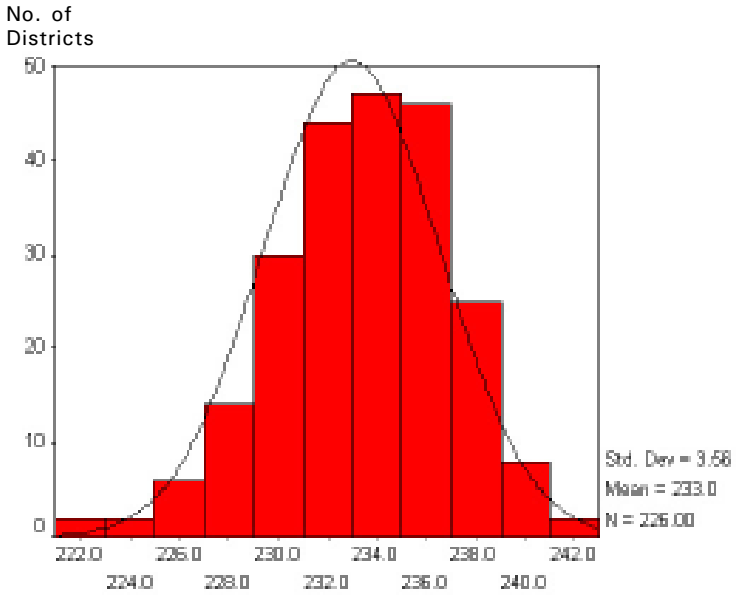
2. Weighted



tests for 1998, 1999, and 2000 were light years away from a normal distribution, and the number of high scores was curiously low. (Compare Figure 5, which shows the distribution of scores on the fourth grade English test for 1999.) Critics who suggested that this test was too difficult appear to have had a good argument.[30]

Figure 5

Fourth Grade English Test, 1999



We now turn to our central concern, which is to trace the sources of variation in district scores as fully as we can. We do this by estimating a spatial regression equation, which we present formally in Appendix 4. This equation allows us to disentangle the factors that affect district MCAS scores and to explain our statistical results in plain English. We also discuss the rationales for testing certain variables.

We began by testing basic economic and demographic variables that are widely thought to affect school performance.[31] Almost every study, for example, finds a relation between a school or district's test scores and its average income. Many also suggest that other economic variables, such as the poverty rate, make a substantial difference. To the extent that it makes any sense to distinguish these from "economic" influences, many studies also indicate that certain demographic variables are relevant. We examined a very large number of economic and demographic variables. Along with average income, these included the percentage of the population receiving assistance under the Temporary Aid To Families With Dependent Children Program (TAFDC); the percentage of families with two parents living in the house; the percentage of families in which one or both parents attended college; poverty rates; the percentage of students classified in DOE statistical reports as white, African-American, Asian, Hispanic, and so forth.[32]

Like almost everyone else, we found many of these variables exceedingly difficult to distinguish from one another statistically. In several cases we relied on so-called "principal components" regression; in others we examined how the respective variables worked in various combinations, along with standard indicators of multicollinearity such as Condition Indices. Such efforts can easily produce the

statistical counterpart of "cabin fever": The addition of one or another variable sometimes promises (or threatens) to reshape the whole equation. Census and DOE "racial" classifications, for example, proved fertile ground for all sorts of complications, since "white" and many "minority" categories can cancel or re-enforce each other, depending on the sample and the facts of the case. Some variables also prove to be related to school dropout rates. On the other hand, "limited English," a category that one might have expected to be closely related to some Census classifications (such as Hispanic), turned out not to be strongly correlated with many independent variables.[33]

Eventually we settled on three economic and demographic variables that give the best overall fit and, in our dataset, are not multicollinear. These are the district's average income, the percentage of two-parent households, and the percentage of the population receiving TAFDC payments. Our statistical results, presented in equation form in Appendix 4, suggest that every $1,000 increase in a school district's per capita income raises average MCAS scores by approximately half a point. A one percent rise in the percentage of families with two parents, by contrast, raises scores about one-seventh of a point. (Here it may be wise to recall that among large groups, apparently small differences in scores frequently signal impressive discontinuities.) The extent of poverty within districts appears to have a major effect on MCAS scores: For every percentage point increase in the number of recipients of aid through the program for Transitional Aid to Families with Dependent Children, MCAS scores decline over a point and a half.

Like others who have pursued such inquiries, we are impressed with the importance of these variables, which are, in a fundamental sense to be discussed later, extraneous to schools. While one or another study might argue for another closely related variable that we ultimately rejected, such as percentage of families with college educated parents (too highly correlated with income), or a district's poverty level (we rely instead on TAFDC percentages), our results here probably will not surprise most other analysts.

Our examination of DOE "racial" variables produced a somewhat more complicated result. To begin with, we note that the usage of "race" here is severely misleading. Like most anthropologists, we reject any idea of pure, biologically definable races. Ever since the first humans left Africa, groups have been intermingling and interbreeding. Neither do we have any truck with hereditary theories. We consider "bell curve" theories of genetically determined intelligence plainly false and absurdly oversimplified.[34] To the extent that categories such as "white" or "Hispanic" make sense, it is as indicators of broad (and commonly, diffuse) "cultural-historical" influences.[35] Language and history in given ecological and material contexts are crucial, not unique group genes.

It was not difficult to find census categories for race that produced plausible initial results.[36] But as our investigation became more specific, these variables — from "Asian" to "Hispanic" and even "white" — receded in importance, with one exception that has an obvious "cultural-historical" interpretation. A one percent increase in the percentage of African-Americans within a district tends to push down a district's average MCAS scores by approximately one-fifth of a point. We caution again that we are looking at district, and not school data, but the origins of this gap are almost certainly not rocket science: Not the biology of the "bell curve," but the legacy of segregation, with its broad implications of chronic underfunding, shaky administration, and weak support structures for high level academic performance.

In recent years, an increasing number of studies have looked beyond demography and economics to assess the influence of health and environmental factors. So did we. Taking advantage of a recently published ranking of Massachusetts towns, we tested for a relation of MCAS scores to environmental hazards within districts.[37] When the equation included our controls for per capita income and other variables, no statistically reliable relation surfaced.

We also considered a range of political variables, which conventional analyses of school achievement and testing rarely examine. With a nod to the old witticism that Alfred Kinsey would have made a memorable contribution to political science if only he had inquired after party identification, we first checked for influence of the partisan affiliations of districts' representatives in the Massachusetts House and Senate. We also investigated whether the statistic that Massachusetts Secretaries of State hasten to report on election night — the percentage of registered voters casting ballots — matters, as well as the much less reported (and far lower) percentage of those voting who were potentially eligible to vote, whether or not they registered. Because we were skeptical that any of these traditional variables would bear much relation to real outcomes, we examined another possibility — whether, in a state like Massachusetts, where so many lavishly financed incumbents face no opponent at all, simply having a challenger on the ballot in 1998 or 2000 made a difference. While assuredly a political variable, this is not one commonly tested by political scientists, who are usually over-impressed by conventional party identifications and regularly confuse "Hail to the Chief" with Aaron Copland's "Fanfare for the Common Man."[38]

Our findings were provocative. The conventional political variables we tested — the party affiliation of district representatives in the Massachusetts House and Senate for 1998 and 2000; voting turnout as a percentage of registered voters, or the percentage of those potentially eligible to vote, registered or not, for 1998, 2000, and averages of both years — all bore no relation to district MCAS scores.

By contrast, in the money-driven Massachusetts political system, where competitive political races are akin to natural wonders, it turns out that MCAS scores of school districts with contested elections to the state Senate in 2000 (by anyone at all, even candidates of "minor" parties such as Ross Perot's "Reform Party"[39]) averaged over half a point higher than elsewhere — to repeat, a considerable amount in a study as highly aggregated as ours.

This result is certainly not conventional in the literature on educational statistics. A number of qualifications are in order, together with some speculations about its meaning. First, it would not be surprising if this finding ultimately proved to be an artifact of some other, more fundamental variable. But the regression controls for the most obvious confounding variable, income. And we have not, so far, been able to find a variable, or set of variables, that takes out the effect.

The basic finding is marvelous in its simplicity: School performance is better in districts where there are contested elections. But the restrictions to Senate elections and the year 2000 raise questions. Elections for the House did not affect district MCAS scores. Neither did facing an opponent in the 1998 Senate elections. The result also holds only when all three years of data are analyzed together rather than individually, though there are perfectly sensible reasons why this should be the case.[40] Our data contain little that can help resolve these doubts, but several facts may be worth noting. First, by 2000, after three years of MCAS testing, educational reform was a front-burner political issue. The Senate was a key political arena. Sen-

ate President Thomas Birmingham, one of the architects of the original reform, championed the issue and was protecting program funding. The House, by contrast, has been less actively identified with the question. Second, 2000 was a presidential election year. As a consequence a great deal more money and organizational support poured into state politics that year. It may be that these extra resources spilled over into some marginally non-competitive Senate districts and empowered challengers, so that there is nothing really mysterious about the unique significance of that election. But real answers require more research. This paper can merely point to evidence of something like a "Birmingham effect" on MCAS scores during the 2000 Senate elections.

We next shifted focus, to search for "school-related" variables that might influence test scores. Once again, we considered a large number. We analyzed the influence of teacher salaries in two ways. First, we examined the impact of maximum salaries in these highly unionized systems. Then we tested for the effect of average teacher salaries. The DOE does not collect any data on superintendents' salaries. But such data are obviously relevant to educational performance. Since superintendents' salaries are public information, we undertook a major effort to collect them for every district in the state. Eventually, after much resistance, we succeeded.

Since, despite recent challenges, conventional wisdom in education studies is that "money doesn't matter" for school achievement, we begin by observing that our results show clearly that it does matter in certain definite ways. Both teachers' maximum salaries and superintendent salaries materially affect test scores: Our equation suggests that raising the maximum teacher's salary (a level presumably enjoyed by relatively few teachers, since it represents the top of a unionized system) by $1,000 drives up MCAS scores by approximately a seventh of a point, while raising the superintendent's salary by $1,000 pushes up district MCAS scores rather less — by about a twentieth of a point. While the latter may not sound like much, one needs to bear in mind that we are talking about a single person's salary, though it would not be surprising if that figure turned out to be related to some other salaries. But not to those of the teachers: Rather to our surprise, the degree of multicollinearity between these two statistics is not damagingly high.

Because we are convinced that a weakness of contemporary quantitative educational research is its lack of interest in managerial variables other than unionization, we experimented with various models of the relation between top management and teachers. But multicollinearity frustrated all of these attempts.

We were surprised to discover that average teacher salaries as opposed to maximum teacher salaries do not appear to matter. It may well be, however, that in highly unionized systems like those of Massachusetts, this statistic mostly reflects the balance struck in any given year between relatively high salaried retirements and new hires who are paid relatively less. If this is the case, it is worth noting that replacing older with younger teachers, assuming that is what lower average salaries signal, did not raise MCAS scores. Nothing in our data supports facile notions that replacing ("burned out") veteran teachers with an influx of eager young pedagogues somehow improves student performance.

For each school district, the state reports how total spending is allocated among various budget categories. Though some DOE data clearly are problematic, officials we queried generally respect the accuracy of these figures, so we subjected them to analysis. We found two that affected scores. One needs only to be described for its

plausibility to be apparent: As total spending rises as a percent of the state's estimate of a district's foundation budget, MCAS scores go up. (The spending figures were for 1997–98, which is important. A primary aim of the state's school financing reforms was to bring every school district up to 100% of foundation spending by 2000. Districts are also allowed to spend more than the foundation budget, and many do.) Our results imply that for every percentage point district spending increases as a percent of the state's foundation budget target, average scores rise by three hundredths of a point — a substantial impact, since the discussion is about percentages that, in affluent districts, frequently exceed 100% of the state target.

By contrast, the other budgetary category that appears to influence MCAS scores operates less conventionally. As spending on books and equipment rises by one percent, MCAS scores actually decline by a hundreth of a point. We find this less mysterious than might be thought. Equipment may be a category in school system budgets where all sorts of expenditures are buried with minimal fanfare. Some of these are surely very large — computer expenditures, for example — and a considerable number may well be wasted.[41] It is also, alas, indisputable that certain districts have sometimes purchased large numbers of the textbooks that they subsequently failed to distribute. Our results, however, do not appear to depend critically on one or two large districts whose problems in this area are well known or widely suspected.

Our investigation of the effects of athletic spending on MCAS performance were quite provocative. They should give pause to anyone who believes that school policies or money makes no difference to school performance. Our regression indicates that for each one percent that a district spends of the amount targeted for athletics in the state's benchmark "foundation budget" for that district, MCAS scores fall by approximately a hundredth of a point. To understand the implications of this result, it helps to remember that some districts spend more than 300% of the (athletic spending) target, dragging down their MCAS scores proportionately.

Why should athletic spending have such negative effects on MCAS performance? The most natural suspicion, that districts that spend proportionately more on athletics are frittering away resources better spent on improving academic performance may well be true, but there is no reason to limit the effect this narrowly. Debates about the role of sports in schooling typically focus on high schools, where the most time, money, and publicity are spent on athletics. But our regression results refer to districts as a whole. We do not find this at all implausible: like analysts who increasingly model schools as "nested" variables, our observation is that school systems are hierarchies, with the high schools normally representing the biggest investment, carrying higher salaries, and setting the tone for districts as a whole. Frequently, lower level sports programs are structured to function as "feeders" for the high school, projecting the dominant level's influence into the lower grades. And, as anyone who has ever lived in a district with a championship football or basketball team can testify, the team's aura radiates far beyond the high school.

This last observation, which is rooted in common experience, and not our data, raises a caution about how the distorting effect of sports might operate in practice. For, after all, it is the district as a whole that suffers: No matter how pathologically averse to hitting the books a whole football or basketball team may be, by itself it cannot drag down an entire district. There are just too few cases.

For sports success to materially lower district MCAS scores, it must affect many more students than those on the teams. Our guess — and it is no more than that — is

that sports success distorts the managerial and academic focus of schools as much as the budget. One of us attended a high school in the Midwest that was a perennial challenger in state basketball tournaments. As tournament time approached, a sort of magical cloud descended on students, teachers, and, at length, even members of the administration. Not only money, but time and energy of both students and faculty were diverted, sometimes for weeks. And all through the year "school spirit" and the sense of achievement it embodied were hard to distinguish from the ghost of championships past. But figuring out precisely how sports success affects MCAS scores is, again, a problem for future research.[42]

Wary of the (handsomely subsidized) flourish of trumpets that accompanies contemporary discussions of schools, computers, and the internet, we checked whether the number of students per computer or the percentage of classrooms with internet access made any difference to districts' MCAS results. We found that the number of students per computer had no effect on test scores. But to our surprise, we did discover that a one percent rise in the percentage of students with access to the web raises test scores by a hundredth of a point. (Once again, since the results are about percentages, this effect is not negligible.) It is possible – we simply lack the data to tell – that this result truly registers the effect of some other unobserved variable. (Note that our equation does control for income, thus we are not simply dealing with an effect of higher incomes.) Alert school managements, for example, may simply be quick to jump on the latest bandwagon, so that we are really measuring managerial quality. The result might even reflect some totally extraneous factor, such as the age of school buildings, since new buildings are now wired for computers, or perhaps the novelty of the internet increases student attention. Once again, however, we cannot make judgments on these issues on the basis of the data in hand.

We wondered if the existence of formal early childhood programs raised test scores or if a district's participation in the METCO busing program might, as some have feared, deleteriously affect performance. We were also anxious to see if variations in student/teacher ratios or the percentage of special education ("handicapped") students mattered. Here our results were mixed. We found no evidence that a district's participation in METCO affected test scores one way or the other. Nor do variations in student/teacher ratios appear to make any difference – though here a simple explanation completely reverses the obvious implication: This ratio does not now vary widely between districts, since it has long been an object of discussion, controversy, and special targeting. While many fear that spiraling costs of special education take resources away from the rest of the curriculum, that battle appears to be over (or perhaps, to have been fought to a draw): We found no evidence that the percentage of special education students affects district test results. On the other hand, when the number of students classified as having "limited English" skills increases by one percent, MCAS scores decline by a tenth of a point.[44]

Another negative result calls for more comment. Nothing we found suggested that early childhood programs had any effect on MCAS scores. Reflecting that most of these programs may be of recent vintage, and suspicious that their effects might wash out as students progressed to higher grades, we substituted fourth grade (grand) averages for consolidated results from all grade levels as the dependent variable. We still failed to find any relation. Perhaps there really is none. But after reviewing these programs, we believe another, more dismal explanation is more plausible. The early childhood variable we used reflects district participation in special

programs targeted for districts with monumental social problems. In the great sweep of American social welfare policy from Reagan to Bush and Clinton, to Bush II, these programs may now be simply too small to make much difference in aggregates as large as school districts.

We, ourselves, believe that we need to be able to measure the alignment of the curriculum with MCAS tests. Alas, the only statewide data known to us is a recent compilation of graduation requirements in school districts that indicate which districts require the study of foreign languages and which do not, how many years of math or English are required for a diploma, and so forth. But it is easy to entertain doubts about the amount of real information contained in this data: The city of Boston, which has some of the lowest average test scores in the state (though with an upward trend), also has some of the toughest requirements on paper. Nevertheless, we coded the data on requirements and analyzed their impact on MCAS scores, but we could not find any effects. Whether four or three or only two years of mathematics are required to get a high school diploma appears to be far less important than the content and quality of instruction across districts.

Finally we examined one of the hottest controversies now raging in American education: whether "competition" from charter or private schools has any demonstrable effect on public school performance. At first impression, our data, which refer exclusively to regular public schools (and thus include no charter schools) might not appear very well adapted for this task. But in fact our data should be highly relevant. Massachusetts law allows towns to permit transfers among schools within a district. A fair number of municipalities have availed themselves of this option. If strong claims about the importance of "choice" in school reform are true, such "intra-district" choice should have some positive effect on MCAS scores, other things being equal.[45] Certain school districts have also taken advantage of a 1991 law to allow students from other districts to enroll or let their own students leave for another district. These cases should also show distinctive MCAS results. (As districts gain or lose students, state funds transit with them.) Finally, formal statistics on the percentage of students who attend private and parochial schools in various school districts do not exist, but a good proxy is surely the percentage of students within a district who do attend public schools. That number is collected by DOE. Accordingly, we investigated the relation of competition to MCAS performance. To our surprise, we found no effect.

Since we do not doubt that either most school boards or major teacher unions are less than enamored of potential rivals, we tried various specifications of all these models, including dummy variables, as well as measures of actual financial losses and gains from transfers of students. The surprise was that none produced evidence that any form of "competition" has thus far affected district MCAS scores.

We suspect that ardent advocates of "school choice" will argue that these findings do not tell decisively against their views. On the whole, we agree, as will become evident below. Still our results do suggest some caveats to fashionable opinion. For example, the failure of intra-district choice to make any detectable difference in MCAS scores perhaps signals that critics of school choice are right in suggesting that the issue is money as much as it is choice. Affording people in poor school districts the option of choosing among inadequate schools does not change anything. Effective demand for good schools, in the sense of demand backed by enough money, simply is not there. Under "choice" plans currently in force, all that may happen is

that desperate parents shuffle their children back and forth between marginal, failing schools. In better districts, the incremental difference in performance may simply not be worth all the transaction costs of abandoning neighborhood schools. The legal possibility of moving, in other words, might be a poor indicator of practical abilities to do so. It may also be that the lack of effect results from uniform conditions imposed by reluctant unions or bureaucratic school boards. At any rate, intra-district choice does not appear to be associated with higher MCAS scores.

Inter-district choice appears to make no difference, either — even in districts that are losing money because students are draining out of them to neighboring districts. Here, the argument can be made that the state has so heavily hedged the possibility of loss, that districts can brush aside their losses indefinitely. This argument is not foolish – there is no doubt at all that the state does indeed hold down the amount of money districts can gain or lose in this way. But there are enough districts losing substantial sums to make one wonder about this hedge.

Our final test of the "choice" view is shadowed by a certain ambiguity, though no way of making the argument will save it. A high percentage of students attending public schools might signal that the schools are relatively good compared to private and parochial alternatives, perhaps because they draw support from the whole community. Or it may indicate that few viable alternatives exist, though in Massachusetts Catholic schools, at least, are common. But there is also a sense in which a low percentage of students in public schools points to the existence of some form of effective competition. Whatever meaning one attaches to the percentage of students enrolled in public schools within districts, though, it appears that the figure is unrelated to MCAS test scores, in the sense that adding the variable to our model yields no increase in predictive power.

III. Seeking Value Added: What Income May Not Buy

At Borodino no one, not even Napoleon, could discern who was winning or losing: "Sometimes cries could be heard through the firing; but it was impossible to tell what was being done there." At no time does the MCAS debate more strongly resemble Borodino than when one attempts to distinguish which districts have done particularly well or poorly. Many of the special problems of analyzing score changes over time are discussed in the next section of this paper. But many disputes about who is running the best race arise from smoldering suspicions that the contest is rigged from the start. As one prize-winning columnist for the *Boston Globe* eloquently expressed these misgivings in the context of the argument over high stakes tests:

> That argument would be more compelling if MCAS did not tell us what we already knew: Poor, urban schools perform much worse than resource-rich suburban schools. Maybe Wellesley kids are test-savvy, or maybe their scores have more to do with better equipped libraries; higher-paid teachers; smaller classes; safe, comfortable homes; and high-achieving parents.[46]

High stakes testing distorts school priorities.[47] The crucial question for public policy is whether these distortions do less harm than the loss of information and accountability that such tests provide. This is a question beyond the scope of this

paper. But a compelling response is possible to the main objection leveled in this passage.

The response begins by agreeing with its central premise. Students in affluent districts do indeed begin with many advantages that help sharpen skills that MCAS measures. (Note that MCAS purports to measure skills, in sharp contrast to other kinds of tests that claim to measure the rather more nebulous notion of "aptitude."[48]) That is why variables such as income, the percentage of TAFDC recipients, and the percentage of two parent families, show prominently in our cross-sectional equation.

It is also easy to understand the associated beliefs that "demography is destiny" and that schools themselves are powerless, or nearly so, to enhance fundamental student skill. But the usual econometric evidence adduced in support of these stronger claims overstates a good case. We have already noted that the typical approach to studying school district performance under MCAS bypasses the problem of spatial autocorrelation; there is thus good reason to be suspicious of conventional assessments of the relative importance of various factors in school performance. With spatial regressions, however, partialing out the influence of particular variables is somewhat trickier, since there is no single easily computed and widely accepted criterion analogous to the famous R^2 in linear regression that can serve as a gold standard. Nevertheless, we agree that all reasonable assessments will end up giving heavy weight to economic and demographic factors.

But our results indicate that customary statistical evidence is problematic: Key school and policy-related variables are often omitted. Until our paper, for example, no one appears to have tested whether superintendent salaries mattered for MCAS performance. Equations purporting to predict MCAS performance left it out. The same holds for other variables that our analysis suggests may have important effects, such as internet access, the extent to which school budgets fell short of foundation budgets in the early stages of reform, and, possibly, even the state of district political competition. (Almost everyone considered the influence of teacher salaries, though not the specification we find to be most powerful, teacher maximum salaries.) The evidence on athletic spending is particularly thought provoking, since our equations indicate it is so large.[49]

The truth is that we simply do not know enough about what influences human learning to tell how far the state of Massachusetts can move sensibly and incrementally to counterbalance social and economic factors that affect school performance. But we can be fairly sure that steps like raising maximum teacher and superintendent salaries would enhance school performance even if we cannot hope anytime soon fully to unlock the deeper mysteries of income, ethnicity, and social class.

There is another way to approach this whole question, which bypasses the usual debates: Use the economic, social, and demographic variables that we have shown predict high performance on the test to control for the advantages enjoyed by affluent districts. Then see which districts do better — discover, that is, the districts that consistently outperform their economics and demography.

In essence, this is an approach to school district performance in terms of *value added.* The general methodology for approaching the question is well known: One drops all school-level variables and regresses only the economic and demographic variables in the original equation on the grand average scores. What results is the test score that these variables alone would predict for each district. By comparing predicted with actual test scores (to get a so-called "residual" score), we can quickly identify which districts are doing better or worse than their demography.[50]

Here, once again, our earlier demonstration that MCAS scores are spatially autocorrelated becomes important. If the original equation is misspecified, so will be any derived from it. Thus our regression equation for "value added" presented in Appendix 5, is, again, a spatial regression.

Rankings for individual school districts in terms of "value added" appear in Appendix 5. Some of these are fascinating. Here, at last, is a list of high performing schools whose entries cannot possibly be dismissed as simple reflections of socio-economic advantage. Still near the top are some affluent, high scoring districts whose performance has been touted in the media on the basis of raw scores. But worries that so-called "ceiling effects" might crimp progress in the state's highest scoring districts turn out to be groundless: The bottom of the list contains some stunning surprises, including several of the most affluent school districts in the state.[51] In the next section of our analysis we examine whether improvements in school performance as measured by MCAS can be explained by the state's sweeping reforms in school financing in the nineties.

IV. The Question of Change: More Than Pennies for Their Thoughts

At Borodino, Tolstoy claimed, fighting raged over three separate stretches of ground. But, the onetime cavalry officer noted, the two side actions were "detached and of little importance in comparison with what took place in the center of the battlefield." The same is not quite true for MCAS, since questions of what drives district MCAS scores and value added are certainly pivotal. Yet, even before the latest state budget crisis, the central analytical question about educational reform in Massachusetts was surely whether the state's colossal investments in K-12 education since 1993 have had any demonstrable affect on school district performance.

Researchers from several different econometric traditions have pursued this question with imaginative research designs. No one, however, has so far reported any unambigously positive results. One analyst who tried an unusual "regression continuity" approach reported evidence that the influx of new funds since 1993 raised scores on MCAS's "low stakes" predecessor, the Massachusetts Education Assessment Program (MEAP). But he detected an effect only for fourth grade tests and only in historically low spending districts.[52] All other grades registered no effects. Analysts who have scrutinized data for MCAS itself all report mixed or even negative effects. Customarily, they temper their findings with the caution that it is too soon to tell, because analyses of changes in test scores are uniquely complex and difficult.

Their diffidence is not a smokescreen. The fundamental problem arises from a stark statistical fact — one that is appreciated by scarcely anyone but specialists: Test scores contain an astonishingly large random component — according to some quite careful estimates, as much as forty percent of all the year to year variance in school scores could be illusory.[53] Sorting out real changes from purely accidental fluctuations is much more difficult than most politicians, educators, or citizens are likely to believe. And how to do so has been highly contentious, even among statisticians.

Consider the Nixon-like career of the statistic most people would instinctively identify as the obvious starting point for analyses of change — the raw "gain" or "change" score. This is calculated as the difference between successive administrations of the "same" (really, comparable) tests. A generation ago, nearly all studies

considered change scores useful evidence. Then came a series of sharp critiques, as statisticians took a closer look at its properties. Soon a variety of successively more complicated substitutes proliferated: "residualized change scores," "estimated true scores," "forward estimated true scores," and so forth. Just as suddenly, the pendulum of professional opinion swung back. Critiques of the critiques blossomed. Change scores began staging what two celebrated statisticians recently hailed as an "amazing comeback."[54]

Some of the thorniest issues derive from the phenomenon of "regression to the mean." This is almost diabolically subtle: Because of random error in scores, high scoring schools frequently do slightly less well the second time round, while low scoring schools tend to score higher. The resulting illusions would be comic, if their consequences were not so incendiary: principals and superintendents in the high scoring schools tear out their hair over the apparent decline in performance — or failure to meet DOE standards for "improvement." Simultaneously their counterparts in low scoring schools bask in community approval and campaign for "rewards" commensurate with their proud "improvements." In neither case, however, has any real change taken place. All one may be witnessing is another round in a dice game. Or the battle of Borodino, where Tolstoy insisted, generals could not easily tell where the lines were or who was winning.

The problem is particularly nettlesome in small schools, where low numbers of pupils taking exams can permit random effects to bulk correspondingly larger. There is no doubt MCAS scores are affected. Both statisticians and journalists have noticed the frequency with which small schools bounce on and off rosters of the state's "most improved" schools.[55]

Additional misgivings arise from the implications of the crucial fact, noted earlier, that social scientists only rarely can perform controlled experiments. Normally, they have to make do with data they find. [56] These have led to a realization that relying on only two data points frequently nourishes "spurious negative correlations" because the "measurement errors in the pre-test (the first of two "scores" under analysis) and the observed change score are negatively correlated." [57] As a result, studies of change that rely on only two data points have now become rare, with the use of multi-wave panel data the norm.[58]

Such statistical obstacles help cast a shadow over the effort to link money to school outcomes. Many who have studied the question claim that "money doesn't matter," at least not very much. They argue that public schools — particularly unionized public schools — are inherently inefficient. While some of these critics believe they can demonstrate that good teachers can indeed raise students' scores, they maintain that school bureaucracies and teacher unions dampen incentives for good teachers. Though some research now claims to qualify these views or find important exceptions, this negative appraisal still dominates the literature.[59]

Our approach to this battleground borrows a leaf from commanders on both sides at Borodino, who were acutely conscious of the need for planning their lines of attack. As before, we first try to clarify the nature of the data. In this instance, a few pictures are worth thousands of words. In the spirit of recently developed techniques for analyzing panel data (which is technically what our sample of repeatedly measured districts is) over time, we plot every district's scores for all three years in which the tests were clearly comparable. The result is a set of so-called "growth curves" for the individual districts. But there are 226 of these, so we aggregate them

into four overarching patterns according to simple rules. This distillation summarizes much more efficiently the broad pattern of change among MCAS scores over three years.

The critical step comes next. We analyze how a wide variety of social and economic variables shape these growth curves through the application of recent techniques of "multi-level" or "hierarchical linear" modeling. Emphasizing the importance of always relying on more than two measurements in any analysis of change, this highly flexible approach treats the time pattern of individual district scores as a first level of analysis. Potential explanatory variables become a second level. Both levels are estimated at once in a single equation with a unified set of error terms. Once again, the details of the equation and its somewhat complex estimation procedures appear in Appendix 7.[60]

As noted above (Figure 2), on average, district scores declined from 1998 to 1999, before rising in 2000. But experiences were not uniform. Figure 6 plots grand averages for every district for each year and connects the scores.[61] The resulting set of individual district "growth curves" is dizzyingly complex.

Figure 6

Three Year Growth Curves — All Districts

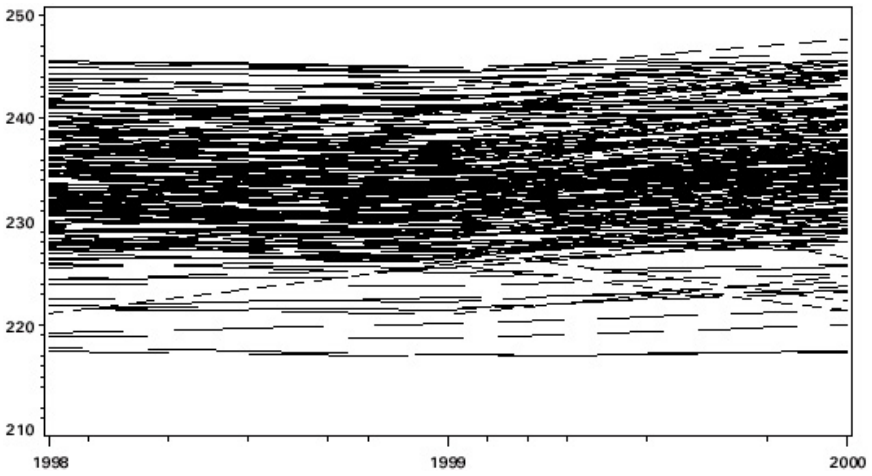
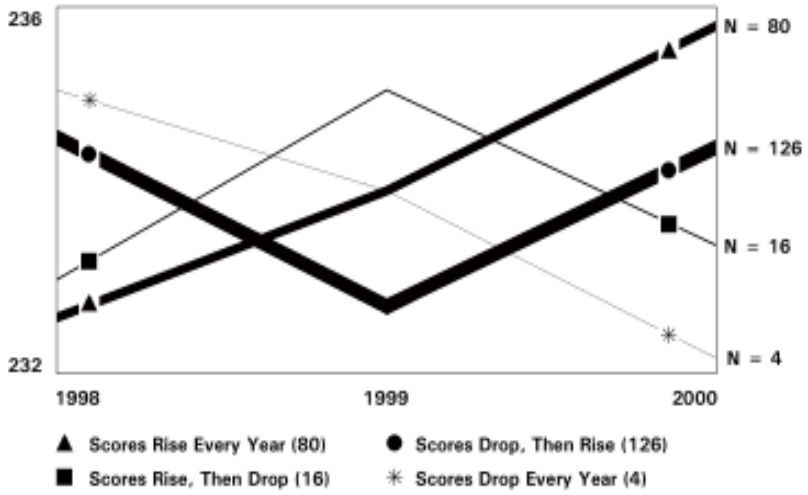


Figure 7 represents our effort to abstract out basic patterns of change in the raw scores. We sort every district into one of four exhaustive and exclusive categories: The handful of districts whose scores went up with each test; the four that declined each time; those where scores first fell and then rose (the most common pattern); and finally the relatively few districts whose scores first rose and then fell. Provided one bears in mind that each of the thick lines represents the mean of a different distribution, around which the districts belonging to each group are scattered, the figure can be treated as a summary of performance in the state as a whole.

To investigate the relation between these changing district scores and the state's investment in K-12 education, the original figure with 226 growth curves has to be

Figure 7

Change in District Grand Averages



translated back into sets of numbers and then analyzed in relation to changes in state spending in each district. Here recent advances in the multilevel modeling of growth curves provide real assistance. First, we use all the data to compute an average rate of change for each district. Where the district scores are plotted as a function of time, this has an easy geometric interpretation: The average annual rate of change is equal to the average of the slopes of each year's change in score with respect to time (one year, in both cases). Appendix 6 displays numerical values of each district's rate of improvement.[62] It is apparent that these rates vary. Now the question is whether these rates of improvement can be explained by the state's sharply rising investments in K-12 education, along with perhaps other district characteristics.

Here it is important to understand how the 1993 Massachusetts Education Reform Act altered school financing. As in other states that introduced similar reforms around that time, a major intent of the legislation was to raise the level of spending in poorer districts that had historically spent less. The idea was to supplement traditional, locally based sources of funds — principally the property tax — with substantial amounts of direct state aid along with a required minimum level of local tax contributions. The vast influx of new money was also clearly intended as part of a grand bargain that would induce school boards and teachers unions to accept what became MCAS and some structural reforms in the way school services are delivered. In this respect, the Act aimed at improving education in the state as a whole.

The scale of the subsequent financial transfers is indeed impressive, though it is important to remember that, because of the colossal spending cutbacks the state made during the earlier recession, for the first few years most of these funds went to restore the status quo.[63] The main channel for state aid has been so-called "Chapter 70" funds, which flow to school districts under a variety of complex formulas. The state established a target "foundation budget" for every district, which varies with district characteristics and is adjusted every year to take account of changing circum-

stances. The foundation budget is not a design for living — it merely sets a floor for expenditures. Districts can, and many do, spend far more, especially, it appears, on athletics. Initially many poorer districts spent much less than their target foundation budgets in many categories. Under the reform, state aid increased in successive years, so that by 2000 every district was at least spending at the minimum (foundation) level.[64]

The MCAS scores we analyze in this paper fall in this period of transition; and our cross-sectional regression indicated that the amount of money that a district spent under or over its foundation budget definitely affected its scores. There is thus good reason to believe that bringing every district up to at least 100% of foundation spending would indeed influence school district performance. But this is quite a different question from showing that changes in MCAS scores in general are systematically related to changes in state aid.

Many factors could muddle results, even if a relation really exists. For example, though direct aid to school districts from the federal government is generally very modest, some districts in the state — mostly poorer ones — do receive significant amounts of money. To the extent this money is wisely used, it will loosen the correlation of score change and *state* aid. Affluent districts are also free to increase spending from their own resources and, as the recent history of school override votes illustrates, some certainly do. This, too, will lower any correlation, as might the less commonly discussed possibility that some districts could take advantage of the state aid to restrict the growth of their own contributions.

A particularly tricky question is precisely how long it takes spending to affect district MCAS results. We have data for actual total spending for every district from 1994 and per pupil expenditure on regular day students from 1993 onward. Data also exists on the amount of "Chapter 70" spending for each district. It is possible to imagine many plausible models of how these might influence district scores. One might in particular wonder how important current spending is compared to past spending, particularly as one considers scores from higher grades, where the students are older and in school longer. If past spending has lagged effects — either because it takes some years to cumulate to something meaningful, or because spending in the more remote past eventually fades into insignificance — it could be difficult to detect even a fairly strong relationship.

We considered many possibilities, including models that closely resemble equations often estimated by econometricians. But our data are technically repeated measures on the same districts. In testing for the effects of money on MCAS scores, it is essential to take account of this data "dependence," or one will end up reporting far more independent observations that there actually are. For this type of problem, so-called "random coefficient" or "multilevel" models provide a superior approach and are now widely used for such tasks within educational statistics.

For technical reasons such models have to be estimated by complex iterative methods — often, but not always, by maximum likelihood. Thus one is well-advised to search for relatively straightforward models.[65] Considering the variety of ways state money might affect MCAS scores, we were surprised to find a simple model that was also consistent with our cross-sectional findings. Again, Appendix 7 presents the full equation. Here our concern is with the meaning of the results.

The model estimates how (linear versions of) the 226 "growth curves" of Figure 6 vary and the relation of this variation to different explanatory variables. Both the

starting point and the slope of each line provide important information — the first about factors that influence the initial level of district school achievement; the second about what controls the rate of progress. One thus uses the multilevel approach to model both the intercept — where the line starts — and the slope, analyzing each separately in relation to different explanatory variables.

In light of our cross-sectional findings, discussed earlier, our results here are striking. Seven variables that figured prominently in the cross-sectional model influence the intercept. Analogously to the way each worked in the earlier equation, four — per capita income, the percentage of families with two parent families, teacher maximum salaries, and superintendents' salaries — operate to raise it. Three — the percentages of households receiving TAFCD assistance and proportionately higher numbers of ESL and African-American students — push it down.[66]

By contrast, we have so far identified only two variables influencing the slope, the rate at which districts improve from their initial position: The amount by which state Chapter 70 aid rose between 1994 and 2000 as a percent of actual total district spending and the change in household income within the district between 1989 and 1999. Since both of these are percentages and can be easily confused with other, similar sounding numbers, it is worth taking a closer look at each.

The figure for change in median household income is a straight measure of percentage change, obtained by subtracting median household income in each district for 1989 from the corresponding data in newly released 1999 U.S. Census figures and then dividing by the 1989 figure.[67] On average district median household income increased by about six percent. But the now widely discussed polarization of American income distribution holds also for Massachusetts school districts: The range of variation was enormous.[68] A fair number of districts actually registered negative growth rates in median household income, while others showed increases as high as 75%. Our statistical results, presented in Appendix 7, suggest that for each one percent that median household income increased between 1989 and 1999, district MCAS scores rose annually by just over a hundredth of a point between 1998 and 2000. (To put this in perspective, it is helpful to recall that on average, scores rose by approximately 1.76 points a year, though, once again, the range of variation was huge — running all the way from 2.8 to a negative 3.14).

The change in state aid is calculated as a difference of two percentages. The first is the percentage of district actual total spending state aid represented in 1994; the second is the same ratio for 2000. We subtract the former from the latter and employ the difference in our equation to estimate the effect of state aid. In almost all districts, this difference is positive. The average increase is about 7% though the range of variation runs from low negative rates to 34%. Our statistical inquiry, presented formally in Appendix 7, indicates that for each 1% this ratio crept up, MCAS scores rose each year by approximately two hundredth (.02) of a point.[69] No current ethnic or racial variables, including the percentage of African-American students in districts, appear to affect slopes of district-level improvements, nor did the absolute level of per capita income.[70]

Since these results have large implications, some amplifying comments are in order. First, just as before, we spent considerable time exploring alternative specifications. We checked to see if high percentages of other minority groups in districts, such as Hispanics, or students with limited English, influenced the intercept and, of course, whether ethnic or racial variables systematically affected the slope. Just as in

the earlier case (and repeating all our earlier cautions about the potential pitfalls of multicollinearity and levels of analysis[71]), we conclude that our data indicate they do not — with one qualification. As discussed earlier, our data for ethnicity and race come from data for schools released by the DOE. We do not have data for longer run changes in these categories on par with that recently released by the U.S. Census for towns. There are good reasons to be cautious about switching from data for schools to town data, but we are currently exploring the use of data from the 2000 U.S. Census to study longer run demographic changes in districts. It is possible that some demographic change in districts that affects scores may come to light as we address this problem, though since our equation takes account of changes in income, it probably catches the most prominent source of variation. We checked to see if high percentages of other minority groups in districts, such as Hispanics, or other variables influenced either the intercept or the slope. Because some other studies have claimed to find impacts, we did employ data from the 2000 Census to test whether changes in the percentages of individuals in poverty or single parent families between 1989 and 1999 affected changes in MCAS scores. They did not.[72]

In any case, it is important to note that by comparison with the vast differences in district starting points, the (ameliorative) effects of state aid have thus far been quite modest. As discussed in more detail in Appendix 7, a substantial amount of our equation's ability to predict change comes from specification of the intercepts. On the other hand, in sharp contrast to previous studies, there is no doubt that the term for state spending, along with the change in household income, significantly improves the equation. And, of course, the effect applies broadly, to district grand averages, rather than simply to one grade level.

The most challenging finding, however, was that if one attempted to predict either the level of 1998 MCAS scores alone or the rate of improvement from 1998 to 1999 with an appropriately truncated version of the same state spending variable, there was no effect. Indeed, on the most plausible specification, the effects of state spending up to that point were negative.

After scrutinizing the data, we are confident that our findings for the entire period are solid. They appear to reflect a real evolution in the state's system for financing education. In the last two years, the state substantially increased the flow of funds to many more districts. The signs of modest positive effects appear when one takes account of this data. The gradual emergence of these effects as state investment widened and deepened may help explain why earlier analysts had such trouble finding their traces — there were few, if any, to be found then. It also suggests that the relation between spending and improvement is still evolving, perhaps rapidly. A particularly intriguing implication is that rhetoric about "staying the course" has a foundation in reality: For district performances to rise in response to the change over six years in the percentage of Chapter 70 funds as a share of actual district spending suggests that districts react to what they perceive as credible long run financial developments, not year to year financial flows.

But, of course, the result also raises other interesting questions: Did the blaze of publicity that accompanied the advent of MCAS play any role in this process? Did the mounting furor over the possibility that large numbers of students might eventually be denied a diploma move districts to more effective action? Did school boards and superintendents perhaps not only have more money to spend, but also attend more rigorously to how they spent their new resources once district scores went up

in the lights? Or are the critics of MCAS right to fear that schools are simply learning to teach to the test?

For now, this latter seems an oversimplification. It is difficult, for example, to think of any reason why districts should learn to teach to the test in proportion to the shift in the amount of Chapter 70 aid they received over six years. No less importantly, Massachusetts scores also rose on the National Assessment of Educational Progress (NAEP), which is widely regarded as the best existing educational assessment vehicle. It is true, however, that in other states where scores on "high stakes" tests have risen through time, such as Texas, national test scores sometimes have not.[73] Particularly given the suspiciously large rise in MCAS scores in 2001, it will be important to benchmark MCAS results against the NAEP data. If the NAEP data begin to lag behind the MCAS results, then alarm bells should ring across the state.[74]

Conclusion: Securing Reform

Almost from the moment the final shots died away, the argument started over who really won the battle of Borodino. Not surprisingly, Tolstoy, like many other Russians, was greatly impressed by the fact that "Napoleon's generals, Davoust, Ney, and Marat, who were close to the region of fire, and sometimes even rode into it, several times led immense masses of orderly troops into that region. But instead of what had invariably happened in all their previous battles, instead of hearing that the enemy were in flight, the disciplined masses of troops came back in undisciplined, panic-stricken crowds. They formed them in good order again, but their number was steadily dwindling." In fact, however, it was the Russians who finally had to withdraw, after suffering over forty thousand casualties.

The struggle for education reform in Massachusetts is going much more smoothly than Napoleon's invasion of Russia. Our analysis of changes in MCAS scores – which all pre-date the recent change in test scoring and thus are proof against challenge on those grounds – indicated that school district performance is truly improving and that the Commonwealth's massive investments in K-12 education are at least a small part of the reason why. Our cross-sectional analysis identified specific school and district practices that appear to be related to superior performance, such as higher superintendent salaries, higher teacher maximum salaries, and limits on athletic budgets.[75] Given the state's continuing commitment to funding reform, there is no obvious reason why such policies could not be more widely emulated. Indeed, we found evidence that the great, cumulative disparities in income and other advantages that demonstrably influence where districts start from might not pose insuperable barriers to improvement even in heavily disadvantaged and minority districts. Our efforts to measure the "value added" by district school systems also show how, if there is a will to do so, MCAS data can be employed in far reaching ways that do not automatically buttress the popular conclusion that the richest schools are the best schools. Indeed, our estimates of value added indicate that some of the most affluent districts in the state rank at the very bottom.

But the battle of Borodino still holds a powerful lesson for education policy in the state of Massachusetts. With his supply lines overstretched and his army exhausted, the French Emperor found himself unable to properly follow up as the tide appeared to turn in his favor at the end of the day. This allowed the Russian army to escape, ensuring that, though the battle was won, the war was "definitively lost."[76]

Massachusetts is now in the midst of a major fiscal transition. Slower economic growth, the bursting of the stock market bubble, and a multi-stage tax cutting initiative are all combining to squeeze state revenues. At the same time the federal education act just signed into law mandates much wider (and more expensive) testing. Napoleonic insight is not required to see that the state's education reforms will inevitably come under increasing pressure at precisely the moment the evidence suggests that the tide has turned.

Greater efficiency or superior information will not solve the problem. But especially as tests multiply under federal pressure, it is absurd for policymakers and the public to rely on summaries of school district scores that are typically in error thanks to easily remediable statistical mistakes. Nor, now that we have shown that the data are spatially autocorrelated, can any public purpose be served by continuing to argue policy in the newspapers or the legislature on the basis of studies that rely on ordinary linear regressions, with no correction for spatial autocorrelation.

The most effective ways to follow up, however, will almost surely emerge from closer analyses of what is being done with all the money and better targeting. Consider, for example, the following sobering fact: Our study indicates clearly that higher maximum salaries for teachers and superintendents enhance district performance. With all the attention and money lavished on K-12 education in recent years, one might, accordingly, suspect that salaries would have exploded. But this does not appear to be true. We lack good time series data on either maximum teachers' or superintendents' salaries. But the Department of Education has published data on average teachers' salaries by district. These are not ideal for constructing precise estimates of statewide average teacher salaries, but they can be used to provide an estimate of how average salaries have changed. The data indicate that average salaries rose by about $1000 a year between 1993 and 1999. Using our estimated average salary for 1993 of approximately $37,415 as the base, this works out to an increase of about 2.75% a year — not much more than the rate of inflation and about half the rate at which average pay within the state rose over the same period. [77] When one examines the correlation between the increasing state aid received by districts in this period and changes in average teacher salaries, a remarkable fact emerges: Changes in school district Chapter 70 spending between 1993 and 1999 and changes in average teachers salaries over the same period are almost uncorrelated.[78] Whatever districts spent their new monies on, it mostly was not higher teacher salaries.[79]

It is easy to reel off other examples where intelligent targeting might make a big difference and perhaps actually save money, by reducing wasted expenditures. In developing our cross-sectional equation, for example, we examined whether some ways school districts allocate expenditures among various categories, such as spending for teachers' aides, or central office spending, influenced performance on MCAS. In most cases we found no effects on test scores. It is true that a negative result is not necessarily conclusive evidence that a particular form of spending is irrelevant — some statistics, such as student/teacher ratios are thought to be so important that every district watches them closely, and thus little variation results. But we are skeptical that this accounts for the apparent irrelevance of so many types of expenditure. Inevitably, one wonders if district performance would improve if more funds were consciously directed into forms of spending that demonstrably enhance district performance.[80] We suspect the answer is yes, but it is time to find out for

sure. Considering the potential size of the effects, a closer look at spending on athletics is also certainly in order.

Our evidence suggests that the high hopes that animated educational reform in the state are actually being fulfilled, and to an extent that surprised us. But while it is good to have a study that documents this, it would be even better if a variety of other studies confirmed this judgment and perhaps amplified or qualified it. It is particularly important to analyze data below the level of districts. Data at the school level are desirable for such a follow up, but there is no reason why Massachusetts, like other states, should not make individual level data, appropriately safeguarded to protect privacy, available to researchers. Such data are particularly important for fashioning better indices of how school performance changes from year to year — a task that the new federally mandated test make more urgent.[81] They are also the best way to check the evidence at the district level that suggests that improvements in scores are not systematically affected by minority status.

In the two years since we began this study, the state Department of Education has greatly improved the amount and timeliness of the data it releases. But this paper's findings in regard to the salaries of school superintendents demonstrate that for all the progress that has been made, we remain uncomfortably close to the situation of the commanders at Borodino. We know the earth is shaking and vast movements are in progress, but it is very difficult to discern what is really happening. To be sure that, unlike Napoleon at Borodino, the initial successes of educational reform in Massachusetts are properly pursued, the press, researchers, and the public need access to far more data on both districts and schools.

It is time to require the Department of Education to gather and release on its website in easily downloadable form not only superintendents' salaries, but average and maximum salaries for teachers and principals. The DOE should also have to release its data in a form that permits meaningful evaluations. There is no excuse, for example, for the Department to continue releasing some scores without indicating the number of students who took the test, so that proper weighted averages can be constructed.[82] Given recent evidence about the importance of institutional grading patterns for student learning and performance, the state should also publish data on grading patterns within schools and districts.[83] Data on district investments in teacher development would also be extremely helpful.

Here, however, there appears to be a disjunction between what the public plainly desires and the current policy. In Massachusetts, the normal pattern of "popularization" in the best sense has broken down almost completely in regard to statistical assessment and discussion of MCAS. A newspaper report about some of our findings in regard to athletics and value added revealed many officials with responsibility for setting policy in districts and in the state clearly were unfamiliar with the idea of statistical controls or even the notion of "value added." They seemed not to realize that the aspirins they reached for as their constituents suddenly began inquiring about value added, instead of raw scores, were tested in accord with essentially the same statistical principles. The media also continue to report unweighted MCAS averages, a practice that is uncomfortably reminiscent of *Alice in Wonderland*.

It is doubtful that the state can go on like this very much longer. DOE officials have now begun to recommend that district school boards be held accountable for continued low performance on MCAS. New federal legislation also requires districts with underperforming schools to offer students transportation to better schools. Both

of these practices raise pointed questions about proper standards for assessing school and district performance.

Not every idea that has surfaced is equally promising. There has been talk of concentrating oversight on districts that appear most in need of it and leaving largely alone districts that appear to be performing acceptably. This may sound sensible, but it is not: It almost certainly implies that affluent districts with high raw scores will be awarded "get out of jail free" cards. Poor districts, especially small poor districts, whose scores are most subject to random variation, will get hammered.

This would be doubly unfortunate. The Commonwealth has an obvious interest in picking up performance in affluent districts with low value added. And there is no point in wasting time and resources turning poor school districts upside down if their central problem is that they are poor. To properly discharge its oversight role, the DOE needs to move in the direction of a formal value added standard, probably including measures that take account of score changes over time. The latter, however, will run headlong into all the concerns about randomness earlier discussed. Unlike some critics of MCAS, we do not believe that these issues cannot be rationally approached or practically solved. Other states are already wrestling with these questions and the new Bush education bill mandating sweeping testing programs guarantees that they will soon move to a political front burner. If the state does not do better than it has so far managed in its efforts to identify top performing districts, the reaction is sure to be sharp.[84]

What will assuredly destroy prospects for continued progress is the impression that DOE is working out its standards behind closed doors, with no public debates and no serious discussion in the media. Because the Department is subject to so much mistaken criticism, it is easy to understand why it often appears to adopt a policy of simply ignoring critics.[85] But this is sure to undermine public confidence in the long run. In the case of standards for replacing school boards or closing charter schools, it is also guaranteed to produce a bitter backlash, as losers discover the arbitrariness inherent in poorly formulated standards adopted without widespread public understanding of what they really entail. Think, for example, of what will happen when a school board or charter school discovers that properly weighted averages would show that its performance was actually better than, say, a dozen other school districts – something that, while infrequent, actually occurs in the data of this study. If for no other reason, the certainty of future lawsuits should motivate state and local officials to look long and carefully at the likely consequences of settling on a standard. But there is also a better reason to encourage wider public discussion of the tests: the likely result would be a broad refocusing of attention and debate on what actually works to boost student performance, and consolidation of the progress that our data appear to signal.

A Healey Public Service Grant from the University of Massachusetts Boston provided modest but important research support. The John W. McCormack Institute also assisted with a small grant to defray some costs of data analysis. We are grateful to Daniel Feenberg for early advice and encouragement. Special thanks to Stephen Klein of Rand for comments on a draft. Thanks also to James K. Galbraith and the editors for additional assistance. The views expressed in this essay are those of the authors alone, not those of the university or any organization with which they are affiliated. All correspondence should be directed to Ferguson, thomas.ferguson@umb.edu., with a copy to Chen, jie.chen@umb.edu.

Notes

1. Leo Tolstoy, *War and Peace*, trans. Constance Garnett (New York: Modern Library, 1994). A verst, by the way, is an old Russian measure of distance, about two-thirds of a mile.
2. The MERA was enacted in 1993.
3. Thomas J. Kane, David Dvorin, and Rachel Deyette, "An Update on School Reform in Massachusetts" (Paper presented at the Conference on Education Reform, Gordon Public Policy Center, Brandeis University, March 9, 2000); Robert Gaudet, "Effective School Districts in Massachusetts: A Study of Student Performance on the 1999 MCAS Assessments," (Boston: Donahue Institute, 2000); see also Jordana Hart, "Study Measures Impact of Demographic Factors on School Performance," *Boston Globe*, January 14, 1999, B1.
4. For a very interesting analysis of the types of errors made by students in the Boston school system, see Lisa Gonsalves, "Learning From the MCAS Exam: An Analysis of Boston Public High School Student Responses," University of Massachusetts, Boston, 2001, photocopy.
5. For the link between home prices and school quality assessments, see David Figlio and Maurice Lucas, "What's in a Grade? School Report Cards and House Prices," (Cambridge: National Bureau of Economic Research, 2000). Their data are for Florida, but we have no doubt that a similar study for Massachusetts would produce analogous results.
6. In the wake of a *Boston Globe* column by Scott Lehigh, March 1, 2002, A17, that reported on some of our results, some DOE officials briefly suggested that because the tests were not yet required for graduation, some students may not have taken them seriously. The inference was that rankings based on the results would not be valid, since the students were not putting forth their best efforts. But this objection confuses random with systematic disturbing factors. Only the latter present any problems. Unless one believes that students in certain districts are somehow more susceptible to this problem than others, there is no reason to be particularly concerned. Everyone took the test under the same conditions. There is also the obvious fact that the tests are clearly being taken extremely seriously, indeed so seriously that threatened boycotts by opponents have dwindled almost to nullity.
7. Spatial statistics as a field has developed rapidly, but it is not widely appreciated within general econometrics or statistics. See, e.g., Noel Cressie, *Statistics for Spatial Data*, (Revised ed.; New York: Wiley, 1993) or A.D Cliff and J.K Ord, *Spatial Processes: Models and Applications* (London: Pion Limited, 1981). A recent helpful applied discussion is Daniel A. Griffth and Carl G. Amrhein, *Multivariate Statistical Analysis for Geographers* (Upper Saddle River, N.J.: Prentice Hall, 1997). In the closing stages of our work on this paper, we finally did encounter work on school test scores that treated them spatially. See Edward B. Reeves and Harold Harty, "Regional Disparities in Kentucky Academic Index Scores," (Morehead, Ky: Morehead State University, 1998) and Timothy Pitts and Edward B. Reeves, "A Spatial Analysis of Contextual Effects on Educational Accountability in Kentucky," (Morehead, Kentucky: Morehead State, 1999).
8. There is now a large literature on GIS systems; see, e.g., Paul Longley and Michael Batty, *Spatial Analysis: Modeling in a GIS Environment* (Cambridge: GeoInformation International, 1996).
9. Test administration is evolving rapidly. The changes, however, can be followed conveniently at either the DOE's website or that of the *Boston Globe*; thus we do not attempt to trace them here. It suffices to observe that early public discussions centered on the three tests we analyze here: English, math, science; and that these were administered as discussed. In 1999, the Department of Education introduced a test in history and social science. In 2001, the DOE decided to give the test in that area, along with the science test, to grade 5 pupils. Some English and math tests have also been moved to different grade levels and some tests have become more specifically focused.

10. Compare Gaudet, "1999 MCAS," or Kane, Dvorin, and Deyette, "School Reform."
11. Ambiguity arises because one adds one district by combining those that then disappear into the portmanteau district. Note that many of these districts enroll relatively few children, so that student coverage figures look better than they are. Since as mentioned below, results of the various tests are highly, but imperfectly, correlated with one another, simply dropping a grade level might also raise additional questions about the comparability of aggregate scores.
12. Not all variables require re-weighting. Following the logic of district consolidation, the most reasonable way to register maximum faculty salaries, for example, is simply to record the highest level in the system; the same holds for superintendent's salary, where there is more than one superintendent. In all cases known to us, the district with the high school dominates such accounts. In the lists of districts in the Appendices, the consolidated districts necessarily are denominated by the names of only one of their components. The particular name we chose has no significance; any name worked, as long as there was but one. The consequence is that someone who lives in one of the other component districts can identify his or her district's proper scores by locating the name of the one component that is in our lists. The DOE website has a useful guide to districts where one can readily find the names of the others.
13. This problem has received much discussion within educational statistics since the appearance of the famous "Coleman Report." See, e.g., Christopher Achen, *The Statistical Analysis of Quasi-Experiments* (Berkeley: University of California, 1986). We are not overwhelmed, however, by claims that modeling the assignment problem slips past this conundrum. In many cases, the chances of mistakenly characterizing the assignment process are at least as high as going wrong from incorrectly modeling the original problem.
14. The problem of multicollinearity is discussed in virtually all treatments of regression analysis. A particularly lucid account, with many suggestions for analysis and treatment, is Stanton Glantz and Bryan Slinker, *Primer of Applied Regression & Analysis of Variance* (Revised ed.; New York: McGraw Hill, 2001). Separate tables breaking out scores by race and gender can be found on the Department of Education website. Since the tables can be sorted by district or school, one might reply that a partial correction is built in. It is true that a comparison of, say, Newton with Chelsea, would control for some factors influencing the scores. But this comparison of a very rich district with a very poor one really begs all the important questions, as is obvious from our cross-sectional equation below. Less glaring contrasts will be largely unrevealing.
15. The existing corrections for different price levels appear to be based on wages in individual towns, plus an allowance for areas with below-average wages. The latter is about redistribution, not price correction. See the discussion in "Proposed Changes to Chapter 70," a working draft available on the DOE website at http://finance1.doe.mass.edu/chapter70/c70_proposals.html. We are currently studying price indices based on a wider labor market study that the state also commissioned. Better price indices should improve our model's fit; we expect to publish on this in the near future.
16. Thus we have only a little to say in this paper about how various ethnic groups fare across districts and our discussion below of factors influencing changes in test scores needs to be read with this qualification. To profitably attack this problem, individual level data would be very helpful. For a good general discussion of the problems of this sort, see Christopher Achen and W. Phillips Shively, *Cross-Level Inference* (Chicago: University of Chicago Press, 1995).
17. See, e.g., Anthony Bryk and Stephen Raudenbush, *Hierarchical Linear Models* (London: Sage, 1992) or Tom Snijders and Roel Bosker, *Multilevel Analysis* (London: Sage, 1999).
18. The data on superintendents' salaries are inevitably fairly noisy. Most of our figures came from *The Massachusetts Bay Cooperative Data Study Report* for 1998, published by the CO/OP, Inc., then in Waltham; the rest we gathered directly from districts. Noise in the data comes from two main sources. First, many dis-

tricts pay varying parts of superintendent health insurance premiums and other benefits, including travel allowances, which are not reflected in our figures. Second, though the fact that the lion's share of our data came from one source is somewhat reassuring, absent detailed knowledge about how each district reports the salary, there may be differences in the meaning of "1998" as between fiscal and calendar years. The consistency of our results suggests that such concerns should not be overblown, but they require noting here.

Our analysis of changes in test scores later in this paper is shadowed by the great cost of calculating change over time in school demographic and other data. As we discuss below, some of our data comes from the 1990 U.S. Census. Between 1990 and 2000, one can only make estimates of changes over five or six years. More importantly, however, many of our key school social variables come from published DOE data. While we inputted a full set of this data by hand for one year, doing much more than this calls for more resources than we had. Data on changing district enrollments is available and we did examine it, though without finding much of interest. Our hierarchical model of change also employs data from the 2000 US Census, downloaded from the Boston.com website.

19. See David Figlio and Maurice Lucas, "Do High Grading Standards Affect Student Performance?" (Cambridge: National Bureau of Economic Research, 2000).
20. See, e.g., the discussion in the *Boston Globe*, Feb. 25, 2000, which reports on the controversy occasioned by an article in *Boston Magazine*. The cases involve particular schools, where the exam allegedly either leaked, students were allowed to revise answers, or other improprieties. Several analysts, notably, Stephen Klein and Laura Hamilton, "Large-Scale Testing Current Practices and New Directions," (Santa Monica: Rand, 1999) have suggested that this problem is much more important than commonly realized. As discussed below, we fear their concerns are warranted in regard to the MCAS and that the problem is becoming much larger than policymakers or school officials want to realize.
21. Tolstoy, *War and Peace*, 912. Alistair Horne, *How Far from Austerlitz?* (New York: St. Martin's, 1996) highly esteems Tolstoy's account of Borodino. One might wonder, but it makes no difference for our points here.
22. We have made no use of the tests for History and Social Science in our analyses in this paper, save in Appendix 2, which illustrates how inclusion of the test affects district grand averages in 2001. This is not because we doubt the importance of historical knowledge — on the contrary, we think historical thinking is fundamental to all serious reflection about society and human beings. But the history of this test is complicated and its coverage is still evolving. See, e.g., the *Boston Globe*, July 24, 2001, B1. This quotes a member of the State Board of Education as suggesting that "It's more important to know who Thomas Jefferson was than to know about the Ming Dynasty. It's a question of priorities." In a twelve-year curriculum, it is difficult to believe the choice is really this stark. We cannot help recalling the ringing words of the legendary — and surely apocryphal — campaign speech of the late Senator Hubert Humphrey: "My friends, I tell you that we do not have to choose between jobs and pollution — we can have both!"
23. Aggregating results of separate tests into one overall score is very widespread and deeply engrained in the case of MCAS, as with other tests, among both the general public and scholarly researchers. But the step can be more problematic than usually recognized. We examined correlation matrices for all the tests in a given year to see how closely the various scores moved. A pattern is quite general that can be predicted from the box plots in Figure 3, below: Since fourth grade scores are on average slightly higher, and tenth grade scores lower, with eighth grade scores in between, within grade correlations are extremely high — almost always well over .90. Correlations of eighth grade tests with fourth grade tests (or the tenth grade tests) run slightly lower — usually in the high eighties. Correlations between tests in the most distant grade levels — 4 and 10 — run in the low eighties or seventies. This suggests to us that while the MCAS lies within the empirical bounds of many tests that are actually administered, efforts to study each test and grade level separately might also be rewarding. Cf., for example, the dis-

cussion of math and language tests in Thomas J. Kane and Douglas Staiger, "Improving School Accountability Measures," (Cambridge: National Bureau of Economic Research, 2001). We regard such studies as an obvious next step in research on MCAS.

24. Some students become ill; a few perhaps move between tests. But see below on other, potentially less innocent reasons for this variation, which must surely require careful monitoring in the future.
25. Formally, let Y_{ijt} and n_{ijt}, i = grade 4,8, and 10, j = subject of MCAS of Mathematics, English, and Science, and t = year of taking MCAS, 1998, 1999, 2000, be the test scores and number of students who took MCAS of grade i, subject j, and year t respectively. The *Globe* simply added the test scores as:

$$\text{GLOBE} = \sum_{i=4,8,10} \sum_{j=M,E,S} Y_{i,j}$$

But the grand average of 3 grades, 3 subjects, and over three years is:

$$\text{GRADAV890} = \frac{\sum_{i=4,8,10} \sum_{j=M,E,S} \sum_{k=\ 98,99,2000} Y_{ijk} n_{ijk}}{\sum_{i=4,8,10} \sum_{j=M,E,S} \sum_{k=\ 98,99,2000} n_{ijk}},$$

and average for year t is:

$$\text{GRADAVt} = \frac{\sum_{i=4,8,10} \sum_{j=M,E,S} Y_{ij} n_{ij}}{\sum_{i=4,8,10} \sum_{j=M,E,S} n_{ij}}$$

26. As explained above, we do not make use of the test for History and Social Science. Administration of this test began only in 1999; after that date differences between our results and some published rankings will result not only from the method of calculating grand averages, but whether or not they also include this test. Properly weighted averages for the newly released 2002 MCAS scores are available on the John W. McCormack Institute of Public Affairs website at www.mccormack.umb.edu.
27. Jordana Hart, "In Harvard, Success Follows Success," *Boston Globe*, December 8, 1999, C1.
28. For a description of the MCAS data reported by the DOE, see the Department's *The Massachusetts Comprehensive Assessment System: Guide to Interpreting the 2000 MCAS Reports for Schools and Districts*, available on the DOE website.
29. Richardson cites an unpublished report by Theodore Micceri, entitled "The Unicorn, the Normal Curve, and Other Improbable Creatures." This examined distributions of scores on over 400 tests given in "schools, universities, and workplaces across the United States." None of the distributions were normally shaped. Cf., Ken Richardson, *The Making of Intelligence* (New York: Columbia University Press, 2000), 35.
30. What one finally thinks on this question will turn at least in part on one's attitude toward so-called "normed" vs. "criterion" tests. The latter purport to assess based on some external standard; the former compare students among themselves. Newspaper reports of high failure rates in tests of historical knowledge or English grammar often turn out to involve criterion tests, in which students are assessed according to some standard that is presumed appropriate.
31. The premise of this paper is obviously that school districts target high scores. Given the focus in the media and the spillover effects on housing prices and other features of districts, we are convinced this premise is sensible. But we did consider the possibility that districts might try to minimize the number of outright

failures. We stopped worrying about this when we discovered that the correlation between rises in grand averages and declines in failures is a near perfect negative 98.5%. (The correlation is of course negative, since the rise of one implies the other's fall.)

32. Nearly all of our cross-sectional data on school districts, including the ethnic composition of enrolled students, came from the bound version of the Department of Education's *School District Profiles, Vol. 1: 1997-1998 School Year and Vol. 2: Foundation Budget Spending Comparisons Fiscal Year 1997*. The Department of Education has since made available a great deal of additional data on its website. We also use some of this data selectively, since it is usually the only available source of time series data on, for example, average teachers' salaries. (The variable, maximum teachers' salaries, which turns out to be more useful, is as far as we know only available from the bound volume, which does not, however, record average salaries.) Where it was necessary to weight district data we made a point of using statistics (such as enrollments) from the bound volume as far as possible, so that our cross-sectional data are internally consistent.

    Political data came from the relevant State of Massachusetts websites (that for the General Court [the Legislature] and the Secretary of State). Some school districts are located in more than one legislative district; we took account of this by turning variables such as "Democrat" or "Republican" into continuous variables, so that a district could be fractionally one or the other. Other demographic variables not included in the DOE bound volumes, such as the percentage of two parent families, were calculated from U.S. Census data, mostly for 1990, downloaded from the Mass CHIP website. Figures for recipients of Transitional Assistance are state figures for 1998, also from that website. We turned the raw numbers into percentages using MISER town population estimates for 1997.
33. One inevitably asks about the reliability of all this data. Our view is that despite some publicized lapses, most is of reasonably high quality. As indicated below, we have some questions about the "limited English" category. It may be worth observing, however, that this variable always needs to be tested in teasing out multicollinearity problems; often adding it with an ethnic variable produces a result suggesting that what appears to be centrally a demographic question is really a language issue. While these issues are always associated, only confusion results if they are confounded.
34. The literature is huge; we do not believe evidence about gaps in any school tests in one country is very relevant to the general question. For that, one needs cross-cultural and over time evidence; see e.g., Gary Collier, *Social Origins of Mental Ability* (New York: Wiley, 1994).
35. For the use of "cultural-historical" here, see, e.g., Laboratory of Comparative Human Cognition, ed., *Culture and Cognitive Development*, 4 ed., Vol. I, *Handbook of Child Psychology* (New York: Wiley, 1983) or James V. Wertsch, *Voices of the Mind: A Sociocultural Approach to Mediated Action* (Cambridge: Harvard Univesity Press, 1991).
36. As noted above, we examined DOE statistical reports on actual school populations; we did not rely on U.S. Census data for the population as a whole in this part of our investigation.
37. Daniel Farber and Eric Krieg, "Unequal Exposure to Ecological Hazards: Environmental Injustices in the Commonwealth of Massachusetts," (Boston: Northeastern University Philanthropy and Environmental Justice Research Project, 2001).
38. See Thomas Ferguson, *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-Driven Political Systems* (Chicago: University of Chicago Press, 1995), especially the Appendix, for a discussion of weaknesses in conventional accounts of party competition.
39. We did dismiss a few cases where an apparent crank candidacy garnered scarcely any votes; though in theory these could present difficult cases of judgment, empirically they are easily distinguished from serious minor party candidates.
40. It may well mean that sample size matters; and it is not too surprising if 1998

and 1999 bear weak relations to an election in 2000, given the negative results for the 1998 election.

41. Construction expenditures are not included in this data. They are separately budgeted. A former school board member who read a draft of this study commented to us that, indeed, his board had in fact used this category as a general reserve fund. No question of fraud was involved; but the funds were employed where they were needed. Certainly any accounting system will require some such reserve system whether or not there are provisions for them.
42. A former school administrator who read a draft of this study commented that some districts may spend still more on sports than is reflected in the budget data we use. He instances two possibilities: user fees and arrangements by which towns might allow school athletic programs to keep all or part of the gate from major spectator sports. Neither he nor we are clear about whether such expenditures would show in the state data we rely upon. Note that user fees would be strongly correlated with town income; the impact of the other avenue is harder to assess. These points underscore our remarks made below, that further research is imperative before anything is done. One might respond that the one general who indubitably bested Napoleon ascribed his victory to the playing fields of Eton. Any assessment of the role of sports in school will certainly want to consider this type of issue. But it is also important to know if tumescent sports budgets are correlated with other phenomena, such as larger differences between genders on math scores. We are currently investigating this possibility.

    Consider also the potential implications of our findings for the debate over charter schools. Our observation is that few of these have major sports programs. Other things being equal (which they surely are not), this might well facilitate a focus on academics that leads to higher test scores. One might also wonder if many issues about unionization that the national literature on school performance has concentrated on would not look very different if a major, previously unmeasured variable received due consideration. But more of this another time.
43. This is not a guess; it is obvious in the data's restricted range of variation. School boards are also clear that this ratio is watched closely.
44. It does not appear that districts apply a uniform standard in classifying students as possessing limited English skills. This question deserves much more attention than it has so far received. Some data we examined, indeed, raised questions about whether results for some districts could be skewed by such variations.
45. See Susan Aud, "Competition in Education: A 1999 Update of School Choice in Massachusetts," (Boston: Pioneer Institute, 1999); we drew on both DOE data and data from Aud's study (which also relied on DOE data) in our analysis here.
46. Eileen McNamara, "Hard Sell Fails Test," *Boston Globe*, February 28, 2001, B1.
47. See the discussion in Klein and Hamilton, "Large-Scale Testing."
48. The test claims to assess skills, not aptitude or intelligence. Though the distinction is popular, it may amount to less than is often claimed. One reviewer of this manuscript commented to us that he believes that the evidence supports the contention that students who do well on the one type of test also score highly on the other. Claims that MCAS measures achieved levels of skill have been challenged recently. See Walt Haney, "Lake Woebeguaranteed: Misuse of Test Scores in Massachusetts, Part I," *Education Policy Analysis Archives* 10: 24 (2002). This is available on the internet at http://epaa.asu.edu/epaa/v10n24/. The argument is somewhat difficult to resolve, since the test also attempts to sort students into various categories, such as "proficient," etc. But it would help if the Department of Education would reply to serious criticism, instead of simply trying to ignore critics, which must inevitably fan suspicions that it is not quite up to the argument. See the discussion below.
49. It may be that some effects that critics attribute to teachers unions or school bureaucracy are in fact consequences of pressures for big athletic programs. None of the studies we have seen that advance such arguments contain measures of the weight accorded to athletics by schools.

50. Appendix 5 also includes the number of exams taken by students; this is not to be confused with the absolute number of students, since most students take more than one exam. The larger the number, the more confidence one can have in the results. We also computed the residual using a Bayesian approach. This yields virtually the same results. We are currently at work on a paper comparing the two approaches to school test score comparisons in detail and so will not develop this theme further here.
51. "Ceiling effects" (if they exist) reflect the fact that the highest scoring districts are thought to be pushing the envelope of available techniques, so that additional large gains in scores would be unlikely. But it is instructive to compare districts such as Wayland, Wellesley, Weston, and Dover-Sherborn in our Table 5. These very affluent districts share many economic and demographic characteristics, and they are geographically adjacent. Yet their rankings in regard to value added differ substantially. They are not all walking on top of an invisible "ceiling" on the frontiers of achievement. Here is a classic instance where an absence of statistical controls leads to major misjudgments in everyday perception.
52. Jonathan Guryan, "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts," (Cambridge: National Bureau of Economic Research, 2001).
53. Kane and Staiger, "Measures."
54. Donald T. Campbell and David A. Kenny, *A Primer on Regression Artifacts* (New York: Guilford Press, 1999). On the long controversy, see also Linda Collins and John Horn, *Best Methods for the Analysis of Change* (Washington: American Psychological Association, 1991) and Frederic Lord, "Elementary Models for Measuring Change," in *Problems in Measuring Change*, ed. Chester Harris (Madison: University of Wisconsin Press, 1963). A very interesting treatment of many issues is David Rogosa, "Myths of Longitudinal Research," www.stanford.edu/class351/longit2k/myths.txt.
55. See Kane and Staiger, "Measures." Cf. also the Boston Globe's characterization of schools that showed big improvements in 1999: "a hodgepodge of tiny towns, charter schools, and small cities sprinkled across the state." All of these entities are much smaller than most school districts. The article appeared on December 8, 1999, front page.
56. See, in particular, Stephen W. Raudenbush, "Comparing Personal Trajectories and Drawing Causal Inferences From Longitudinal Data," in *Annual Review of Psychology*, 2001, ed. S. T. Fiske, D. L. Schachter, and C. Zahn-Waxler (Palo Alto: Annual Reviews, 2001), 516-523.
57. Stephen Raudenbush and Anthony S. Bryk, *Hierarchical Linear Models* (2nd ed.; London: Sage Publications, 2002), 166.
58. As we go to press, the Beacon Hill Institute (BHI) at Suffolk University has issued a report, "Getting Less for More: Lessons in Massachusetts Education Reform," in which data from only two points in time are central to most of the analysis. Several of its claims require careful scrutiny.

    We have two fundamental methodological objections. The Department of Education plainly warned that various changes in scoring and other features rendered MCAS results for 2001 incomparable with those for 1998, 1999, and 2000. Ignoring recent literature on measuring change, the BHI makes no use of the three years of data that are widely agreed to be comparable; instead, it takes the suspiciously high 2001 data as one of its two comparison points. The other time point it chose does not come from MCAS data at all - the report plucks it from the old, low-stakes Massachusetts Educational Assessment Program (MEAP) for 1994.

    This is not a happy choice. In the terminology of Raudenbush, discussed above, the fundamental problem is to get as good a fix as possible on the "trajectories" of the various school districts as they implement educational reform. It is obvious that the reliability of the trajectories one estimates will be greatly affected by how many data points one employs: The more, the better, always assuming the points reflect comparable measures. Estimates from two points cannot possibly be as reliable as efforts that make use as much data as possible. Omitting the three

most reliably comparable data points and using an entirely different test as a baseline can only engender confusion.

There is a substantial literature on equating different tests; see e.g., Robert L. Linn, "Linking Results of Distinct Assessments," *Applied Measurement in Education* 6 (1), 83-102. The conditions for doing so are quite demanding. One needs to be sure that the same constructs are being measured, with the same reliability. The low stakes MEAP was a shorter test (and thus inevitably less reliable), which did not report results for individual students. It did not reflect a statewide curriculum framework, nor was it accompanied by insistent pressure on ESL and disabled students to participate fully. Considering that statisticians have been unable to agree on a uniform system for equating even well-understood tests such as the SAT and the ACT (see Linn, "Results," 89) because they measure different constructs, it is unlikely that a mix of MEAP and MCAS constitutes a good benchmark. We doubt that even the lower requirement for "calibration," in Linn's sense, is clearly met. The BHI study actually relies upon rankings based on percentiles. These can always be mechanically generated, whether or not comparisons make sense. In any case, correlated errors (in Raudenbush's sense, above), regression to the mean, and other "regression artifacts" described by Campbell and Kenny, may well shadow comparisons from just two data points. The BHI probably does not help itself by foreswearing average scores and sorting outcomes into the categorical boxes in which the DOE sorts students ("Warning/Failure," "Needs Improvement," etc.), and grouping schools by the 1994 test score ("previous performance"). We earlier observed that changes in the DOE category of "warning/failure" track changes in grand averages almost perfectly. Adding boxes almost surely enhances the chances of finding spurious changes, while sorting by track records may easily increase the incidence of what Campbell and Kenny term "regression artifacts due to extreme group selection." For these reasons alone, we would be suspicious of the BHI claims about class sizes, even if our data (for a single point in the nineties) did not suggest that variation in these ratios has been exceedingly limited and, perhaps (as the BHI report notes), mis-measured.

Second, the BHI report suffers from serious - indeed, we suspect, fatal - "omitted variable" problems. It never estimates the impact of ethnic variables, or ESL, though both our equations and common sense indicate that several such factors influence MCAS scores; and it buries athletic spending in broader categories that muffle its impact. In the absence of proper controls, results for the variables the BHI does test are necessarily suspect. The failure to run any tests for spatial autocorrelation inevitably raises additional question marks about many of its estimates.

Happily, our dataset permits a direct test of the BHI study's claim that higher average teacher salaries lower MCAS scores. In our change equation described in Appendix 7, average teacher salaries become a Level 2 variable. We have run the test and its results are not significant (though, in contrast to the BHI's claim, the term is positive, not negative). This is hardly a surprise - as discussed previously, districts have imperfect control over average teacher salaries, which are probably a function of the balance of incoming and retiring teachers, as much as collective bargaining settlements. By contrast, districts have more direct control over maximum teachers' salaries, and this variable shows powerfully in several of our equations. But the BHI did not think to test this variable and we do not have data on changes in maximum salaries over the nineties to run a test ourselves. As discussed below, it is also clear that most districts have not raised teacher salaries that much. Changes in district household income and state aid are far more telling in influencing changes in MCAS scores.

We have other doubts about the BHI report: Nowhere does it attempt a direct test of the effects of state aid on MCAS scores. Neither does the report provide direct evidence about the importance of charter schools. The only variable the BHI study actually estimates in regard to "choice" is changes in percentages of students in public schools. But as we observed earlier, this is an ambiguous indicator, which may mean either that the school system is very good or that affordable al-

ternatives are sparse. No less importantly, changes in the ratio may have little to do with "choice." For example, financial pressures on Catholic schools can affect the percentage of students in public schools and may reflect district per capita income, declines in religious vocations, and other extraneous factors. We are not uncritical partisans of public schools - one of us once joined with the French consul in an unsuccessful bid to establish a bilingual charter school supervised not only by the DOE, but the French Ministry of Education. But the failure of our own direct tests of the impact of charter schools (described above) makes us skeptical about their significance, at least so far. We are not surprised that, as this essay goes to press, stock in the best known private company in the business of running schools is selling for less than a dollar a share.

Given the problems inherent in assessments of change, we believe it is vital that researchers present full-length cross-sectional equations and derive their basic measures of "value added" from these, as we have. Cross-sectional studies are far more straightforward to interpret and can serve as a valuable check on findings from studies of change. We also believe that multicollinearity is a particular problem in analyzing dropout rates, which is why we have little to say about them in this paper.

59. The literature on this topic is too large to be inventoried here. The debate is perhaps most conveniently followed in articles that are currently available on the web, c.f., David Grissmer, Ana Flanagan, and Stephanie Williamson, "Does Money Matter for Minority and Disadvantaged Children? Assessing the New Empirical Evidence," (Santa Monica: Rand, 1996); Eric Hanushek, "Have We Learned Anything New?" The Rand Study of NAEP Performance, (Stanford: Hoover Institution, 2001).
60. On multi-level models, standard sources include Bryk and Raudenbush, *Hierarchical*; Snijders and Bosker, *Multilevel*; and Harvey Goldstein, *Multilevel Statistical Models* (2nd ed; London: Edward Arnold, 1995). Also very helpful are Ita Kreft and Jan de Leeuw, *Introducing Multilevel Modeling* (London: Sage, 1998); Ronald Heck and Scott Thomas, *An Introduction to Multilevel Modeling Techniques* (Mahwah, N.J.: Lawrence Erlbaum, 2000); and Kelvyn Jones and Craig Duncan, "People and Places: The Multilevel Model as a General Framework for the Quantitative Analysis of Geographical Data," in *Spatial Analysis: Modeling in a GIS Environment*, ed. Paul Longley and Michael Batty (Cambridge: GeoInformation International, 1996).

    Our own discussion draws heavily on John B. Willett, "Measuring Change: What Individual Growth Modeling Buys You," in *Change and Development: Issues of Theory, Method, and Application*, ed. Eric Amsel and K. Ann Renninger (Mahwah, N.J.: Lawrence Erlbaum, 1997), 213-43. See also Margaret Burchinal and Mark Applebaum, "Estimating Individual Development Functions: Methods and Their Assumptions," *Child Development* 62, No. 1 (1991), 23-43. We have also profited from several draft chapters of a work in progress by James Ware, especially "Chapter 3. Single Group Repeated Measures Design."
61. This figure is loosely inspired by John B. Willett, "Questions and Answers in the Measurement of Change," in *Review of Research in Education* 15, ed. Ernst Rothkopf (Washington: American Educational Research Association, 1991), 345-422; and David Rogosa, David Brandt, and Michele Zimowski, "A Growth Curve Approach to the Measurement of Change," Psychological Bulletin 92, No. 3 (1982), 726-48.
62. The Table in the Appendix includes the number exams taken by students. For a particular district, the higher the number, the more confidence one can have in the particular figure.
63. See the discussion in Kane, Dvorin, and Deyette, "School Reform." From approximately 1989 to the mid-nineties, the state was in fact, "getting less for less."
64. Note, however, that our regression results indicate that average scores in districts that spent more than 100% of the state foundation budget were higher.

The implication is that the 100% target is simply a specified minimum, not a guarantee of some peculiarly advantaged status.

65. See in particular the discussion in Kreft and de Leeuw, *Introducing*, 82ff. Note that the first level models are often estimated via an empirical Bayes procedures. In our data none of this matters very much.
66. One near-miss may be of interest. Athletic spending came very close to attaining an acceptable level of statistical significance in defining the intercept (.11). Some analysts, indeed, might have taken it over. We prefer caution, but mention the point.
67. Note that, as usual, the Census figures for towns have to be combined and properly weighted to yield statistics for school districts.
68. For the increasing polarization of incomes, see, e.g., Kevin Phillips, *Wealth and Democracy: A Political History of the American Rich* (New York: Broadway Books, 2002) and the review by Thomas Ferguson, "Following the Money," Washington Post Book World 32, No. 20 (2002), 7.
69. This effect is not huge, as we have several times emphasized, and the change of a percent in budgetary terms is a fair amount of money. But the effect is unambiguously positive and, as explained shortly, there are reasons to suspect that it may be increasing. Note, however, that the large rise in test scores in 2001 certainly reflects other factors in addition to broader state spending.
70. Reckoned on the basis of the figures used in the cross-sectional model above, not changes in any of these percentages over the nineties. We doubt that most of these would matter very much, for reasons discussed immediately below.
71. Especially the level of analysis. We cannot stress too strongly that results at the level of schools could differ from our findings for districts, particularly in regard to the possible influence of social and ethnic variables.
72. The data for Massachusetts towns came from the Massachusetts Institute of Social and Economic Research. We are grateful to Dr. Stephen Coelen, the Director, for making these data available to us. The change in single parent families came moderately close to attaining statistical significance. It did not, however, reach even .15, which is considerably below the levels we have accepted in the rest of our study. Note that other measures of poverty, especially ones internal to the school rather than towns, might give different results.
73. See the discussion in Stephen Klein et al., "What Do Test Scores in Texas Tell Us?" (Santa Monica: Rand, 2000).
74. See Brian Stecher and Laura S. Hamilton, "Putting Theory to the Test," Rand Review 26, No. 1 (Spring 2002), 20-21. Note that MCAS scores did not rise in the second year of the test, contrary to the pattern in other states discussed in this article.
75. If, as we suspect, athletic budgets reflect the underlying values of the school system, then simply capping expenditures would accomplish little.
76. The quotation is from Charles Esdaile, *The Wars of Napoleon* (London: Longman, 1995), 259.
77. Our figures for average teachers' salaries come from the DOE website and, incidentally, appear to be the same as those used by other studies. But state-wide averages calculated from this data are only approximations, since the website does not show the number of teachers. Note, however, that we are mostly interested in the changes between the numbers. *The United States Statistical Abstract, 1996*, 425, reports "average annual pay" for 1993 in Massachusetts was $30,229. The corresponding figure for 1999 is currently on the U.S. Bureau of Labor Statistics website. It is $40,352. It may also be noted that state average pay in 2000 rose to $44,326. We do not have the figures for teachers' salaries for 2000. Note that these figures are averages, and thus are probably pulled up by high values at the top of the income scale. Figures for median state pay, if we had them, might indicate that the raises received by most of the Commonwealth's citizens were more closely aligned with those received by teachers.
78. A simple spatial regression of changes in teachers' salaries between 1994 and 1999 on changes in district Chapter 70 spending over the same period shows a

coefficient of .04 times the change in Chapter 70 spending measured in thousands of dollars. The significance level is .09.

79. The question of a possible salary cap for teachers apparently exercised the Massachusetts House in the early stages of educational reform in Massachusetts. See the discussion in Craig Bolon, "Educational Reform in Massachusetts," August 30, 2000; available on the web at http://www.massparents.org/easternmass/brookline/ed_reform_bolon.htm In light of our results for the effect of teacher maximum salaries on MCAS performance, it is fortunate this proposal was eventually withdrawn. Note that the discussion is about raises. Districts could, and certainly often did, also hire more teachers.
80. Many important categories of spending are too indirectly reflected in existing figures to be tested at all. Certainly there are no figures for such important school functions as after-school programs or arts and music. Recent work on local government responses to budget cutbacks suggests some unsettling possibilities about how school districts might be tempted simply to cut such apparently "peripheral" programs in times of budget stringency. In the absence of better statistical information about how schools respond, outcomes that make little sense in either the long or short run can happen as a result of temporary budget crises. See the sometimes chilling discussion in David Figlio, "What Might School Accountability Do?" *National Bureau of Economic Research Reporter*, Fall 2001.
81. The advantages of tracking the same students over time for purposes of evaluation can scarcely be overestimated. This has been done by other states without compromising privacy. See ibid.
82. On the DOE website, the main MCAS results normally come with the number of students taking the test. But this is not true of, for example, its data for results by gender.
83. See Julian Betts and Jeff Grogger, "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry Level Earnings" (Cambridge: National Bureau of Economic Research, 2000).
84. The lists have obvious problems with randomness. See Anne Wheelock, "School Awards Programs and Accountability in Massachusetts: Misusing MCAS Scores to Assess Quality," available on the web at the Fair Test website. Cf., also, *Boston Globe*, June 26, 2002, B2. We would not, however, agree that with appropriate techniques, MCAS scores cannot yield good evidence of school improvements. Everything depends on how one analyzes the data. It may be that a legislative oversight committee will have to compel DOE to do it right.
85. A case in point, perhaps, are some (not all) of the complaints about individual questions. We are sympathetic to suggestions that existing practice is defective and needs more attention. But it is not necessarily true that a few defective questions invalidate a test. One needs to know how many questions and, in particular, whether everyone is equally likely to go wrong on them or not. The latter issue is much more serious.
86. Cliff and Ord, *Spatial Processes*, 1981.
87. Cressie, *Statistics for Spatial Data*, 1993.
88. Stephen Raudenbush and Anthony Byrk, *Hierarchial Linear Models*, 2nd ed. (London: Sage, 2002), 164.

Appendix 1

MCAS Grand Averages for 1998, 1999, 2000 and Three Year Consolidated Grand Average By District

(Ranked by Three Year Average Performance)

Rank	SCHOOL	1998	1999	2000	98--00
1	HARVARD	245.47	244.93	246.37	245.60
2	CONCORD CARLISLE	244.93	244.44	247.68	245.54
3	WELLESLEY	245.54	245.01	245.60	245.38
4	ACTON BOXBOROUGH	244.68	244.90	246.33	245.32
5	WESTON	245.32	244.57	245.53	245.15
6	LEXINGTON	244.29	244.51	245.10	244.65
7	MEDFIELD	244.31	243.99	244.19	244.16
8	DOVER SHERBORN	243.84	244.04	244.49	244.12
9	WINCHESTER	243.93	242.74	245.22	244.00
10	WAYLAND	242.68	243.58	245.55	243.96
11	BELMONT	243.32	243.00	245.44	243.95
12	NEWTON	243.70	242.85	244.23	243.60
13	NEEDHAM	242.37	241.93	245.27	243.19
14	WESTBOROUGH	241.62	242.01	244.57	242.76
15	LINCOLN SUDBURY	241.96	243.55	242.38	242.64
16	WESTWOOD	242.12	242.19	243.23	242.53
17	SHARON	242.99	242.22	242.16	242.45
19	WESTFORD	240.88	240.64	245.12	242.30
19	NORTHBORO SOUTHBORO	242.38	240.82	243.76	242.30
20	ANDOVER	241.44	241.80	243.61	242.29
21	BEDFORD	240.49	241.12	244.17	241.95
22	COHASSET	240.14	241.12	244.34	241.84
23	LONGMEADOW	240.15	241.05	243.89	241.74
24	HINGHAM	241.12	240.65	243.25	241.70
25	NORTH READING	241.79	240.39	242.26	241.52
26	BROOKLINE	241.44	240.80	242.14	241.47
27	NORWELL	241.28	238.69	244.38	241.44
28	HAMILTON WENHAM	240.92	241.26	241.98	241.37
29	SHREWSBURY	240.48	239.69	243.21	241.18
30	DUXBURY	240.31	240.89	241.79	240.99
31	WACHUSETT REG	239.93	241.19	241.72	240.95
32	READING	241.01	239.49	241.92	240.82
33	GROTON DUNSTABLE	241.02	240.22	240.52	240.58
34	MASCONOMET	239.61	240.84	240.84	240.45
36	MANCHESTER	238.05	240.77	241.37	240.05
36	MEDWAY	239.97	238.97	241.17	240.05
37	SANDWICH	239.74	239.98	240.33	240.02
38	HOPKINTON	239.54	237.12	242.51	239.82
39	LENOX	239.44	239.97	239.89	239.78
40	NASHOBA	239.55	238.58	240.98	239.68
41	BERLIN BOYLSTON	238.58	238.93	240.93	239.51
42	LYNNFIELD	239.15	236.99	242.47	239.48
43	LITTLETON	238.44	238.59	240.96	239.31
44	PENTUCKET REGIONAL	238.27	239.52	239.73	239.20
45	ARLINGTON	238.59	239.48	239.27	239.12
46	HANOVER	238.71	238.07	240.58	239.09
47	FRANKLIN	237.45	238.73	240.78	239.04
49	MARBLEHEAD	238.75	237.61	240.64	238.98
49	SCITUATE	239.28	237.13	240.41	238.98

Rank	SCHOOL	1998	1999	2000	98-00
50	NAUSET	237.85	237.30	241.90	238.97
51	SWAMPSCOTT	236.14	240.64	239.50	238.73
52	WEST BOYLSTON	238.16	236.48	240.53	238.44
54	CHELMSFORD	237.89	236.55	240.44	238.29
54	HADLEY	238.35	237.80	238.79	238.29
55	CANTON	236.55	237.54	240.34	238.13
56	FRONTIER	236.80	236.09	241.21	238.08
57	NATICK	238.52	236.32	239.45	238.07
58	WOBURN	238.06	236.38	239.46	237.98
59	NORTH ANDOVER	237.17	237.73	239.02	237.97
60	HOLLISTON	236.75	238.55	238.26	237.88
61	BRAINTREE	235.60	236.85	240.96	237.78
62	NORWOOD	237.18	237.40	238.11	237.57
63	IPSWICH	237.09	235.84	239.85	237.56
64	STONEHAM	237.40	236.82	238.41	237.53
65	AMHERST PELHAM	237.86	238.73	235.49	237.35
66	MILTON	236.82	235.58	239.37	237.34
67	WALPOLE	237.23	236.80	237.93	237.32
68	GEORGETOWN	236.93	234.85	240.33	237.24
69	EASTON	236.01	236.59	239.14	237.22
70	NEWBURYPORT	236.87	237.24	237.41	237.17
72	EAST LONGMEADOW	236.70	236.20	238.56	237.14
72	HAMPDEN WILBRAHAM	236.73	236.66	238.03	237.14
73	TYNGSBOROUGH	235.62	235.70	239.72	237.11
74	ROCKPORT	236.96	234.71	239.92	237.06
75	KING PHILIP	236.09	237.51	237.34	237.00
76	HATFIELD	236.23	238.70	236.24	236.96
77	GRAFTON	237.03	234.71	239.02	236.90
78	MANSFIELD	235.18	235.75	239.18	236.75
79	MENDON UPTON	237.45	235.39	236.64	236.51
80	FOXBOROUGH	234.50	235.70	238.81	236.38
81	MARSHFIELD	235.59	235.91	237.05	236.18
82	BURLINGTON	235.24	234.69	238.10	236.05
83	MARTHAS VINEYARD	235.64	234.94	237.49	236.04
84	AUBURN	236.52	235.94	235.53	236.00
85	CHATHAM	235.22	234.92	237.58	235.84
86	NORTH MIDDLESEX	235.86	235.08	236.29	235.74
87	ASHLAND	235.65	235.27	236.17	235.70
88	WAKEFIELD	235.04	235.28	236.76	235.68
89	SILVER LAKE	235.20	234.52	236.93	235.57
91	NORTON	235.25	234.00	237.15	235.53
91	WEST BRIDGEWATER	235.38	234.14	237.26	235.53
92	BEVERLY	235.35	233.60	237.59	235.51
93	HOPEDALE	234.13	234.72	237.83	235.45
94	DIGHTON REHOBOTH	234.87	234.62	236.54	235.37
96	DANVERS	234.91	233.14	237.27	235.13
96	SUTTON	232.66	235.34	237.12	235.13
97	NORTH ATTLEBOROUGH	234.33	234.78	235.79	234.99
98	MILLIS	234.72	233.18	237.12	234.89
99	OLD ROCHESTER	234.40	233.57	236.49	234.87
100	HAMPSHIRE	233.51	233.74	236.21	234.74
101	QUABBIN	234.14	234.82	235.03	234.67
102	WILMINGTON	234.28	234.06	235.40	234.59
103	BELCHERTOWN	233.82	233.79	236.11	234.58
105	DEDHAM	233.35	232.77	237.51	234.53
105	NANTUCKET	234.43	231.51	237.86	234.53
106	MOUNT GREYLOCK	235.70	232.93	235.06	234.50

Rank	SCHOOL	1998	1999	2000	98-00
107	MILLBURY	233.73	233.56	236.08	234.44
108	MELROSE	233.10	234.23	235.94	234.40
109	ASHBURNHAM WESTMINSTER	233.76	234.27	235.07	234.39
110	TEWKSBURY	233.64	232.61	236.27	234.24
111	NORTHAMPTON	234.14	233.34	235.32	234.23
113	HARWICH	235.57	233.26	233.81	234.18
113	PLYMOUTH	233.72	233.88	234.95	234.18
114	BILLERICA	234.39	232.56	235.50	234.17
115	LUNENBURG	234.62	232.60	235.25	234.15
116	UXBRIDGE	232.19	233.66	236.45	234.14
117	GRANBY	233.02	232.59	236.43	234.13
119	BRIDGEWATER RAYNHAM	234.49	233.76	234.02	234.09
119	CENTRAL BERKSHIRE	235.97	232.82	233.42	234.09
121	MONSON	232.11	234.64	235.30	234.00
121	PROVINCETOWN	230.77	237.08	234.01	234.00
122	HUDSON	233.23	232.98	235.58	233.94
123	TRITON	232.50	234.63	234.35	233.83
124	DUDLEY CHARLTON REGIONAL	233.49	232.36	235.20	233.76
125	STOUGHTON	232.56	233.57	234.90	233.68
126	LEICESTER	232.00	233.37	235.51	233.67
127	FRAMINGHAM	232.92	233.33	234.61	233.61
128	TANTASQUA	231.91	232.88	235.22	233.47
129	BLACKSTONE MILLVILLE	231.80	233.46	235.00	233.46
130	ABINGTON	231.36	233.05	235.62	233.40
131	WHITMAN HANSON	233.19	233.07	233.68	233.31
132	MILFORD	232.43	231.96	235.46	233.30
133	FALMOUTH	232.13	233.05	234.55	233.24
134	BARNSTABLE	232.63	232.67	234.39	233.22
135	AMESBURY	232.76	230.76	235.39	233.05
136	CLINTON	232.20	232.71	234.04	232.96
137	WATERTOWN	230.98	232.92	234.70	232.88
138	AYER	231.98	232.22	234.28	232.86
139	CARVER	229.95	233.30	235.10	232.78
141	DOUGLAS	231.82	230.63	235.67	232.72
141	DENNIS YARMOUTH	232.91	231.38	233.96	232.72
142	AGAWAM	232.41	231.85	233.84	232.71
143	BELLINGHAM	231.49	233.52	232.63	232.54
144	DRACUT	231.79	232.39	233.31	232.49
145	SEEKONK	231.49	231.84	234.04	232.46
146	LEE	233.45	231.60	232.31	232.43
147	DARTMOUTH	232.09	232.29	232.75	232.38
148	SPENCER EAST BROOKFIELD	232.98	230.82	233.34	232.37
149	EAST BRIDGEWATER	231.95	230.01	234.91	232.35
150	MOHAWK TRAIL	231.64	231.36	233.81	232.22
151	SWANSEA	232.50	229.98	233.93	232.08
153	BOURNE	230.96	230.32	234.99	232.03
153	PEABODY	231.40	231.91	232.76	232.03
154	SOUTHWICK TOLLAND	230.51	230.39	235.35	232.01
155	SOMERSET	229.78	230.75	234.99	231.81
156	MAYNARD	232.47	231.87	231.02	231.77
157	ROCKLAND	230.75	230.69	233.60	231.74
158	QUINCY	230.90	231.21	232.59	231.56
159	SAUGUS	231.44	230.96	232.30	231.55

Rank	SCHOOL	1998	1999	2000	98-00
160	WINTHROP	230.10	231.61	232.95	231.51
161	NORTHBRIDGE	230.46	230.53	233.41	231.48
162	MARLBOROUGH	230.68	231.39	232.20	231.44
163	FREETOWN LAKEVILLE	230.80	230.23	233.09	231.36
164	WEYMOUTH	230.90	230.76	232.08	231.26
165	SOUTH HADLEY	231.27	230.03	232.05	231.13
166	NORTH BROOKFIELD	231.31	230.72	231.08	231.03
167	QUABOAG REGIONAL	231.37	229.65	231.83	230.95
168	PIONEER VALLEY REG.	230.85	229.56	232.34	230.93
169	HULL	230.61	229.06	233.09	230.89
170	METHUEN	230.13	230.25	231.99	230.81
171	GLOUCESTER	230.78	228.86	232.81	230.78
172	RALPH MAHAR	230.01	230.79	231.06	230.64
174	HOLBROOK	227.62	230.49	233.22	230.54
174	MIDDLEBOROUGH	230.76	228.48	232.39	230.54
175	MASHPEE	229.52	228.32	233.35	230.48
176	GATEWAY	229.62	230.82	230.67	230.39
177	BERKSHIRE HILLS	231.46	232.89	226.50	230.32
178	LUDLOW	230.54	230.04	230.32	230.30
179	ADAMS CHESHIRE	230.91	228.64	231.31	230.28
180	LEOMINSTER	229.17	229.78	231.36	230.11
181	FAIRHAVEN	229.73	229.70	231.21	230.09
182	OXFORD	230.06	229.63	230.12	229.92
183	GILL MONTAGUE	229.42	229.00	231.18	229.80
184	WALTHAM	229.11	228.54	230.97	229.51
186	ATTLEBORO	229.15	228.61	230.48	229.43
186	GARDNER	227.94	229.36	231.07	229.43
187	WESTPORT COMMUNITY	229.19	229.26	229.02	229.16
188	EVERETT	228.99	228.70	229.17	228.95
189	WEST SPRINGFIELD	227.95	228.82	229.14	228.63
190	MEDFORD	227.82	227.30	230.77	228.60
191	GREENFIELD	228.42	227.35	230.16	228.59
192	RANDOLPH	227.79	226.20	232.04	228.57
193	EASTHAMPTON	227.18	228.80	229.55	228.51
194	PALMER	228.58	227.20	229.64	228.50
195	WAREHAM	227.35	227.72	230.31	228.48
196	PITTSFIELD	229.23	227.89	228.06	228.39
197	SOUTHERN BERKSHIRE	228.02	226.62	230.39	228.34
198	WESTFIELD	227.92	228.01	229.07	228.33
199	WEBSTER	228.83	227.50	227.84	228.07
200	NORTH ADAMS	228.29	226.40	229.75	228.05
201	NARRAGANSETT	226.92	227.83	228.86	227.87
202	AVON	228.35	225.89	229.25	227.79
203	SOUTHBRIDGE	226.32	226.81	229.89	227.66
204	WARE	226.31	226.76	229.74	227.63
205	MALDEN	227.67	227.26	227.85	227.59
206	ATHOL ROYALSTON	227.40	227.04	228.25	227.57
207	SALEM	226.56	226.53	229.65	227.52
208	WINCHENDON	227.24	228.05	227.12	227.46
209	HAVERHILL	226.86	226.05	227.82	226.92
210	REVERE	225.90	225.51	228.55	226.61
211	SOMERVILLE	225.51	226.35	226.44	226.10
212	CAMBRIDGE	228.60	227.30	222.33	226.07

Rank	SCHOOL	1998	1999	2000	98-00
183	GILL MONTAGUE	229.42	229.00	231.18	229.80
184	WALTHAM	229.11	228.54	230.97	229.51
186	ATTLEBORO	229.15	228.61	230.48	229.43
186	GARDNER	227.94	229.36	231.07	229.43
187	WESTPORT COMMUNITY	229.19	229.26	229.02	229.16
188	EVERETT	228.99	228.70	229.17	228.95
189	WEST SPRINGFIELD	227.95	228.82	229.14	228.63
190	MEDFORD	227.82	227.30	230.77	228.60
191	GREENFIELD	228.42	227.35	230.16	228.59
192	RANDOLPH	227.79	226.20	232.04	228.57
193	EASTHAMPTON	227.18	228.80	229.55	228.51
194	PALMER	228.58	227.20	229.64	228.50
195	WAREHAM	227.35	227.72	230.31	228.48
196	PITTSFIELD	229.23	227.89	228.06	228.39
197	SOUTHERN BERKSHIRE	228.02	226.62	230.39	228.34
198	WESTFIELD	227.92	228.01	229.07	228.33
199	WEBSTER	228.83	227.50	227.84	228.07
200	NORTH ADAMS	228.29	226.40	229.75	228.05
201	NARRAGANSETT	226.92	227.83	228.86	227.87
202	AVON	228.35	225.89	229.25	227.79
203	SOUTHBRIDGE	226.32	226.81	229.89	227.66
204	WARE	226.31	226.76	229.74	227.63
205	MALDEN	227.67	227.26	227.85	227.59
206	ATHOL ROYALSTON	227.40	227.04	228.25	227.57
207	SALEM	226.56	226.53	229.65	227.52
208	WINCHENDON	227.24	228.05	227.12	227.46
209	HAVERHILL	226.86	226.05	227.82	226.92
210	REVERE	225.90	225.51	228.55	226.61
211	SOMERVILLE	225.51	226.35	226.44	226.10
212	CAMBRIDGE	228.60	227.30	222.33	226.07
213	TAUNTON	224.37	225.55	228.28	226.04
214	WORCESTER	225.99	225.04	225.86	225.62
215	FITCHBURG	224.39	224.60	225.60	224.87
216	CHICOPEE	224.03	223.80	224.98	224.26
217	BROCKTON	222.51	222.17	224.25	222.98
218	CHELSEA	221.09	225.80	221.42	222.75
219	FALL RIVER	221.89	222.63	223.30	222.59
220	LYNN	221.68	221.01	224.74	222.50
221	NEW BEDFORD	221.89	221.41	223.77	222.39
222	LOWELL	221.35	221.64	223.16	222.05
223	BOSTON	219.19	220.07	221.41	220.23
224	SPRINGFIELD	218.92	218.66	220.04	219.19
225	LAWRENCE	217.81	217.16	217.53	217.50
226	HOLYOKE	217.49	216.98	217.32	217.27

Appendix 2

2001 MCAS Scores
Grand Averages 2001
Without and With History Test Scores

Rank	School	W/O HIST	WITH HIST
1	WELLESLEY	252.12	250.08
2	LEXINGTON	251.19	249.29
3	CONCORD-CARLISLE	251.13	249.00
4	ACTON-BOXBOROUGH	250.58	247.90
5	WAYLAND	250.45	248.43
6	WINCHESTER	250.30	247.89
7	NEWTON	250.10	248.16
8	WESTON	249.61	248.20
9	WESTWOOD	249.38	247.21
10	DOVER-SHERBORN	249.25	247.34
11	COHASSET	248.99	246.77
12	SHARON	248.74	246.40
13	HARVARD	248.71	246.58
14	LINCOLN-SUDBURY	248.65	246.86
15	NEEDHAM	248.59	246.85
16	WESTFORD	248.45	246.68
17	BELMONT	248.24	246.26
19	WESTBOROUGH	247.86	246.48
19	MEDFIELD	247.81	246.12
20	HINGHAM	247.68	245.96
21	LONGMEADOW	247.56	246.08
22	BEDFORD	247.29	245.51
23	GROTON-DUNSTABLE	247.26	245.52
24	LYNNFIELD	247.13	245.47
25	ANDOVER	246.87	245.43
26	READING	246.75	245.04
27	HAMILTON-WENHAM	246.65	245.11
28	BROOKLINE	246.51	244.63
29	NORTHBORO-SOUTHBORO	246.44	244.64
30	SHREWSBURY	246.30	244.32
31	HANOVER	246.24	244.69
32	DUXBURY	245.78	243.49
33	MANCHESTER ESSEX REGIONAL	245.73	244.40
34	NORWELL	245.72	244.53
36	WACHUSETT	245.66	243.83
36	MARBLEHEAD	245.39	243.90
37	NAUSET	245.32	243.28
38	MASCONOMET	245.25	243.27
39	NORTH READING	245.18	244.04
40	HOPKINTON	245.15	243.42
41	SCITUATE	245.13	243.30
42	NASHOBA	244.97	242.66
43	MEDWAY	244.73	243.52
44	FRANKLIN	244.67	242.68
45	ARLINGTON	244.62	242.77
46	BRAINTREE	244.58	242.59

Rank	School	W/O HIST	WITH HIST
47	CHELMSFORD	244.48	243.04
49	HOLLISTON	244.44	242.70
49	NEWBURYPORT	244.39	241.99
50	SWAMPSCOTT	244.32	242.35
51	EASTON	244.07	242.43
52	SANDWICH	244.02	242.23
54	PENTUCKET	243.93	241.94
54	CANTON	243.87	242.27
55	CHATHAM	243.60	241.34
56	FOXBOROUGH	243.47	241.67
57	FRONTIER	243.30	241.54
58	STONEHAM	243.25	241.56
59	MILTON	243.25	241.27
60	KING PHILIP	243.20	241.96
61	LENOX	243.19	241.80
62	NORWOOD	243.16	241.63
63	DANVERS	243.12	241.32
64	LITTLETON	242.87	241.65
65	BERLIN-BOYLSTON	242.84	240.95
66	MANSFIELD	242.81	241.16
67	NORTH ANDOVER	242.75	240.88
68	MARSHFIELD	242.74	241.10
69	WALPOLE	242.51	240.64
70	MELROSE	242.47	240.98
72	MARTHAS VINEYARD	242.45	240.99
72	NATICK	242.40	241.01
73	HAMPDEN-WILBRAHAM	242.39	240.82
74	IPSWICH	242.32	240.27
75	MOUNT GREYLOCK	242.20	239.81
76	BURLINGTON	242.19	241.04

Rank	School	W/O HIST	WITH HIST
107	CENTRAL BERKSHIRE	239.24	237.45
108	SILVER LAKE	239.16	237.64
109	DEDHAM	239.14	237.61
110	BILLERICA	239.09	237.64
111	WATERTOWN	239.04	236.86
113	TANTASQUA	239.01	237.51
113	WILMINGTON	238.98	237.09
114	AMESBURY	238.96	237.39
115	UXBRIDGE	238.90	237.38
116	TRITON	238.79	236.98
117	GRANBY	238.77	237.65
119	BARNSTABLE	238.64	237.08
119	OLD ROCHESTER	238.62	236.95
121	ASHLAND	238.58	236.76
121	ASHBURNHAM-WESTMINSTER	238.56	236.77
122	ROCKPORT	238.53	237.20
123	PLYMOUTH	238.53	236.81
124	EAST BRIDGEWATER	238.49	237.15
125	SOUTHERN BERKSHIRE	238.48	237.00
126	WINTHROP	238.45	237.25
127	AUBURN	238.41	236.85
128	DENNIS-YARMOUTH	238.41	236.73
129	DUDLEY-CHARLTON REG	238.41	236.99
130	BELCHERTOWN	238.13	236.81
131	FALMOUTH	238.11	236.26
132	QUINCY	238.08	236.86
133	NANTUCKET	238.02	237.16
134	WEYMOUTH	238.02	236.53
135	FRAMINGHAM	237.99	236.44
136	SOMERSET	237.93	236.34

Rank	School	W/O HIST	WITH HIST
77	LUNENBURG	242.04	240.05
78	SUTTON	241.96	240.17
79	WOBURN	241.95	240.27
80	GEORGETOWN	241.94	240.21
81	HADLEY	241.33	240.04
82	AMHERST-PELHAM	241.31	239.44
83	EAST LONGMEADOW	241.12	239.35
84	MILLIS	241.02	239.71
85	PROVINCETOWN	240.82	239.13
86	NORTH ATTLEBOROUGH	240.77	239.21
87	MENDON-UPTON	240.65	238.54
88	HATFIELD	240.65	239.19
89	WAKEFIELD	240.64	239.25
91	HOPEDALE	240.59	238.80
91	GRAFTON	240.54	239.00
92	NORTH MIDDLESEX	240.51	238.47
93	NORTHAMPTON	240.47	238.52
94	DIGHTON-REHOBOTH	240.35	238.73
96	NORTON	240.30	239.26
96	TEWKSBURY	240.25	238.91
97	WEST BOYLSTON	240.14	238.58
98	ABINGTON	240.03	238.11
99	BRIDGEWATER-RAYNHAM	239.98	238.38
100	TYNGSBOROUGH	239.92	237.95
101	STOUGHTON	239.87	238.24
102	BEVERLY	239.76	238.13
103	HAMPSHIRE	239.73	238.82
105	QUABBIN	239.43	238.05
105	WEST BRIDGEWATER	239.41	237.46
106	WHITMAN-HANSON	239.40	237.83

Rank	School	W/O HIST	WITH HIST
137	MILFORD	237.88	236.32
138	DRACUT	237.67	236.15
139	BLACKSTONE-MILLVILLE	237.43	235.71
141	AGAWAM	237.40	236.09
141	HARWICH	237.39	235.81
142	CLINTON	237.37	235.31
143	DOUGLAS	237.18	235.51
144	MAYNARD	237.13	235.74
145	SOUTHWICK-TOLLAND	237.07	235.50
146	ROCKLAND	237.02	235.30
147	FREETOWN-LAKEVILLE	236.97	235.47
148	DARTMOUTH	236.85	235.19
149	WALTHAM	236.83	235.55
150	SAUGUS	236.83	235.21
151	HUDSON	236.83	235.39
153	SEEKONK	236.74	235.43
153	LUDLOW	236.70	235.26
154	NORTHBRIDGE	236.70	235.82
155	AYER	236.69	235.14
156	PEABODY	236.66	235.26
157	LEICESTER	236.66	235.35
158	MONSON	236.60	235.12
159	BOURNE	236.47	234.94
160	MOHAWK TRAIL	236.45	234.83
161	MILLBURY	236.36	234.75
162	QUABOAG REGIONAL	236.31	235.07
163	AVON	236.27	234.04
164	BELLINGHAM	236.25	234.28
165	SOUTH HADLEY	236.18	234.99
166	MARLBOROUGH	236.17	234.29

Rank	School	W/O HIST	WITH HIST
167	BERKSHIRE HILLS	236.09	234.79
168	SPENCER-E BROOKFIELD	236.08	235.07
169	HOLBROOK	235.99	234.35
170	ATTLEBORO	235.90	234.25
171	METHUEN	235.79	234.22
172	MIDDLEBOROUGH	235.78	234.58
174	MASHPEE	235.69	234.77
174	CARVER	235.55	233.97
175	LEE	235.51	233.92
176	GATEWAY	235.30	233.48
177	MEDFORD	235.30	233.89
178	NORTH BROOKFIELD	235.22	233.78
179	HULL	235.05	233.84
180	GLOUCESTER	234.99	233.54
181	PALMER	234.98	233.07
182	LEOMINSTER	234.96	233.73
183	WAREHAM	234.90	233.28
184	SWANSEA	234.84	233.77
186	EVERETT	234.75	233.30
186	WESTPORT	234.72	232.93
187	PIONEER VALLEY	234.49	232.50
188	NARRAGANSETT	234.34	233.27
189	WESTFIELD	234.27	232.92
190	GARDNER	234.18	232.68
191	EASTHAMPTON	234.05	232.77
192	RALPH C MAHAR	233.91	232.33
193	MALDEN	233.62	232.19
194	GILL-MONTAGUE	233.52	231.98
195	WEST SPRINGFIELD	233.52	232.01
196	GREENFIELD	233.47	231.97

Rank	School	W/O HIST	WITH HIST
197	SOMERVILLE	233.46	232.08
198	FAIRHAVEN	233.38	232.05
199	ADAMS-CHESHIRE	233.38	232.21
200	RANDOLPH	232.97	231.56
201	REVERE	232.84	231.61
202	TAUNTON	232.62	231.06
203	CAMBRIDGE	232.51	231.08
204	ATHOL-ROYALSTON	232.50	230.82
205	PITTSFIELD	232.42	231.05
206	OXFORD	232.19	230.58
207	HAVERHILL	231.96	230.62
208	SALEM	231.87	230.47
209	NORTH ADAMS	231.52	230.49
210	WARE	231.36	230.18
211	WINCHENDON	231.28	230.02
212	SOUTHBRIDGE	231.21	229.78
213	BROCKTON	230.92	229.52
214	WORCESTER	229.29	228.13
215	WEBSTER	228.98	227.60
216	LYNN	228.53	227.27
217	CHICOPEE	228.52	227.47
218	NEW BEDFORD	228.29	227.08
219	LOWELL	227.95	226.94
220	FITCHBURG	227.78	226.60
221	BOSTON	227.71	226.72
222	CHELSEA	227.58	226.34
223	FALL RIVER	226.57	225.43
224	SPRINGFIELD	224.68	223.70
225	LAWRENCE	224.19	223.21
226	HOLYOKE	222.23	221.41

Appendix 3
Tests for Spatial Autocorrelation in the MCAS Data

Previous studies of MCAS have not raised any questions about the presence of spatial autocorrelation in the data. But this question is important, since if the data are spatially autocorrelated, the most common statistical technique for analyzing the data — ordinary least squares regression — is likely to yield misleading results.

There are essentially two ways of investigating this possibility. The first and most direct is to run a specific test for this possibility. A variety of such tests exist; the most common is probably the so-called "Moran" test. (For a general survey see Luc Anselin and Rosina Moreao, "Properties of Tests for Spatial Error Components," 2000.) In our case, one would test for spatial autocorrelation on the dependent variable — the three year grand average of MCAS district scores. The second is to estimate our main cross-sectional equation via ordinary least squares and then show that its residuals are spatially autocorrelated.

Both procedures indicate that spatial autocorrelation is present in the MCAS data. Formally, in the former case, where Y_i, $i = 1,2 \cdots 226$ is the grand average of MCAS scores for years 1998, 1999, and 2000 for district i, a Moran test shows that the estimated spatial autocorrelation is .3467 with a test statistic of 8.613 and p-value = 0. We report results for the other test, on the residuals of ordinary least squares versions of our cross-sectional model, in Appendices 4 and 5. These indicate that purely linear models of MCAS suffer from the presence of spatial autocorrelation. By contrast, Moran tests and plots of the residuals of all of our spatial models (omitted from this paper for reasons of space) indicate that they are not autocorrelated.

Appendix 4
Cross-Sectional Spatial Regression Equation for Predicting School District MCAS Performance

Appendix 3 showed that the our dependent variable, the three year grand average of school district MCAS scores, is marked by appreciable spatial autocorrelation. An analysis of an ordinary least squares version of our cross-sectional equation reinforces this conclusion. Formally, let Y_i be the three year grand average of MCAS scores of the i[th] district $i = 1,2,\cdots,226$, and let $X_{ij}, j = 1,2,\cdots 12$ be the independent variables listed below in Table A4.2 of this Appendix.

We test for spatial autocorrelation for the residuals of the ordinary least squares (OLS) version of our cross-sectional model using an S-PLUS function moranForLM.q developed by Kaluzny, based on a formula of Cliff and Ord.[86] Formally the OLS model is :

$$Y_i = \beta_0 + \sum_{j=1}^{12} \beta_j x_{ij} + \varepsilon_i \qquad \text{(A4.1)}$$

The small P-value of the test results indicates that the residuals are indeed spatially autocorrelated. Table A4.1 lists the estimated spatial autocorrelations of the residuals of the OLS model and the associated test statistic.

Table A4.1

Moran's autocorrelation test for the residuals from an ordinary least squares linear regression model. Equation (A4.1):

Moran Autocorrelation	Z statistic	P-value
0.2077	5.5801	0.0000

We now consider a spatial regression model. Once again, let Y_i be the response variable of the three year grand average of MCAS scores for the i[th] district, $i = 1,2, \cdots ,226$. Following Cressie, we now employ a Simultaneous Spatial Autoregression (SAR) model as follows:[87]

$$Y_i = \mu_i + \rho \sum_{j=1}^{n} w_{ij}(Y_j - \mu_i) + \varepsilon_i \qquad \text{(A4.2)}$$

where the error term ε_i is assumed independent and identically normally distributed with $N(0, \sigma_0^2)$; μ_i is the mean effect defined as follows:

$$\mu_i = \beta_0 + \sum_{j=1}^{12} \beta_j X_{ij} \qquad \text{(A4.3)}$$

and $X_{ij}, j=1,2,\cdots 12$ are the covariate variables listed in Table A4.2 below; while W is the neighborhood matrix defined as:

$$W_{ij} = \begin{cases} 1 \text{ if } i \text{ is connected to j} \\ 0 \text{ if } i = j \text{ or if } i \text{ is not connected to j} \end{cases} \quad \text{(A4.4)}$$

In matrix notation, Equation (A4.2) can be written as:

$$Y = X\beta + \rho W(Y - X\beta) + \epsilon \quad \text{(A4.5)}$$

or

$$(I - \rho W)(Y - X\beta) = \epsilon \quad \text{(A4.6)}$$

provided the covariance matrix of Y, $[(I - \rho W)'(I - \rho W)\sigma^2_0]^{-1}$ is symmetric and positively-defined. Both ρ and σ^2_0 are unknown scale parameters that can be estimated by minimizing the negative log likelihood of Equation (A4.6). If ρ is fixed, the m.l. estimators of β are

$$\beta = (X'(I - \rho W)(I - \rho W)X)^{-1}X(I - \rho W)'(I - \rho W)Y \quad \text{(A4.7)}$$

We estimate this using the SLM function of the program S-Plus, as discussed earlier. Table A4.2 shows the estimated coefficients and associated p-values. Figure 1 of this Appendix, below, plots the residuals of Equation (A4.2) as a histogram and in the form of a Q-Q normal plot. In sharp contrast to our results for (A4.1), these indicate that the residuals from the spatial regression model approximate very well a normal distribution. Not surprisingly, when the Moran test is run on this model, the result is a large p-value indicating that the residuals are not spatially autocorrelated.

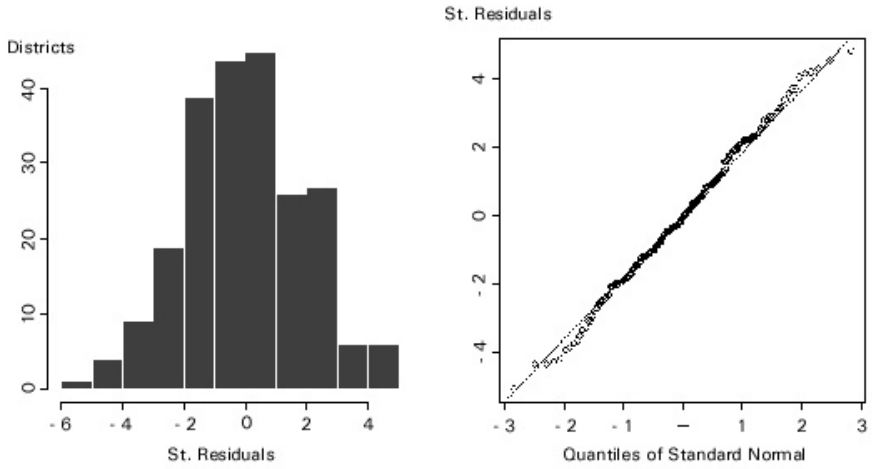
Table A4.2

Column 3 contains the estimated coefficients and P-values estimated with S-PLUS

Variable	Definition	S-PLUS -- Coefficient
TPERCAP	Per Capita Income (Unit: $1000)	.46 (.00)
TWOPHLD	Percent of Households with 2 Parents	.13 (.00)
TAFDCPER	Percentage Receiving TAFDC	-1.60 (.00)
TSAL98S	Superintendent's Salary – 1998-99 (Unit: $1000)	.05 (.00)
AFRICAN-AMERICAN	Percentage of Students of African-American Descent	-0.24 (.00)
TSMX989	Teachers Maximum Salary 1998-99 (Unit: $1000)	.14 (.00)
INTERNET	Percent of Classrooms with Internet Access	.01 (.01)
TOTAL	Total spending as % of Foundation Budget '97	.03 (.01)
LIM.ENG	Percent of Students Classified as Limited English Speakers	-0.10 (.07)
ATHLETIC	Percent of Foundation Budget Spent on Athletics	-0.01 (.03)
BOOKEQUI	Percent of Foundation Budget Devoted to Books and Equipment	-0.01 (.04)
SENATE O	Senator Faced Opponent in 2000 (1 = Yes, 0 = No)	.66 (.07)

Appendix 4, Figure 1

Histogram of residuals based on model A4.2 (Left)
Q-Q Normal Plot of residuals based on model A4.2 (Right)



Appendix 5
School Districts Ranked By Value Added
Residuals For 4 Economic/Demographic Variables
(See Text – Details of Economic/Demographic Equation Below)

RANK	SCHOOL	98-00	RES890	3 YR TOTAL EXAMS
1	AMHERST PELHAM	237.35	6.73	8709
2	LENOX	239.78	6.03	1701
3	HARVARD	245.60	5.57	2451
4	WESTBOROUGH	242.76	5.01	6251
5	BELMONT	243.95	4.61	7083
6	NAUSET	238.97	4.46	6837
7	NORTH READING	241.52	4.42	4468
8	NORTHAMPTON	234.23	4.20	6101
9	ACTON BOXBOROUGH	245.32	4.20	10307
10	HAMILTON WENHAM	241.37	3.89	5026
11	SANDWICH	240.02	3.79	7946
12	ARLINGTON	239.12	3.45	8153
13	NEWTON	243.60	3.34	22600
14	HADLEY	238.29	3.33	1274
15	BEDFORD	241.95	3.31	4113
16	MEDFIELD	244.16	3.13	5382
17	NORTON	235.53	2.96	5351
18	FRONTIER	238.08	2.94	3277
19	SHREWSBURY	241.18	2.92	8239
20	BRAINTREE	237.78	2.86	9794
21	NORWOOD	237.57	2.83	6848
22	WOBURN	237.98	2.74	9113
23	MEDWAY	240.85	2.74	4785
24	READING	240.82	2.67	8300
25	BEVERLY	235.51	2.57	9693
26	BELCHERTOWN	234.58	2.54	4618
27	PENTUCKET REGIONAL	239.20	2.53	6901
28	WEST BOYLSTON	238.44	2.52	2284
29	MILLBURY	234.44	2.51	3542
30	NEEDHAM	243.19	2.43	8797
31	MANSFIELD	236.75	2.43	7361
32	WACHUSETT REG	240.95	2.37	13365
33	WELLESLEY	245.38	2.36	6971
34	SWAMPSCOTT	238.73	2.36	5536
35	BROOKLINE	241.47	2.33	11691
36	STONEHAM	237.53	2.30	5578
37	CHELSEA	222.75	2.23	9175
38	FRANKLIN	239.04	2.22	9321
39	DENNIS YARMOUTH	232.72	2.22	8965
40	HANOVER	239.09	2.21	5112
41	WESTWOOD	242.53	2.18	4769
42	FRAMINGHAM	233.61	2.14	14584
43	SPRINGFIELD	219.19	2.09	46012
44	EAST LONGMEADOW	237.14	2.02	5384
45	PROVINCETOWN	234.00	1.99	856
46	NATICK	238.07	1.96	7916
47	SOUTHBRIDGE	227.66	1.93	4872
48	ANDOVER	242.29	1.91	11851

RANK	SCHOOL	98-00	RES890	3 YR TOTAL EXAMS
49	NEWBURYPORT	237.17	1.76	4954
50	PLYMOUTH	234.18	1.75	17213
51	STOUGHTON	233.68	1.68	8248
52	NORTH ADAMS	228.05	1.59	5160
53	WESTFORD	242.30	1.59	8157
54	MOUNT GREYLOCK	234.50	1.58	3398
55	LEXINGTON	244.65	1.53	11771
56	SHARON	242.45	1.53	6936
57	CONCORD CARLISLE	245.54	1.50	6953
58	AYER	232.86	1.48	3515
59	BARNSTABLE	233.22	1.44	13970
60	CLINTON	232.96	1.42	3809
61	NORTHBORO SOUTHBORO	242.30	1.41	8005
62	ROCKPORT	237.06	1.37	2343
63	WEST BRIDGEWATER	235.53	1.35	2107
64	MARTHAS VINEYARD	236.04	1.35	4892
65	CHATHAM	235.84	1.34	1379
66	AGAWAM	232.71	1.30	8643
67	SCITUATE	238.98	1.30	5997
68	MILFORD	233.30	1.29	7853
69	RALPH MAHAR	230.64	1.23	3686
70	HINGHAM	241.70	1.23	6934
71	MILTON	237.34	1.21	7726
72	TANTASQUA	233.47	1.21	6797
73	FALMOUTH	233.24	1.13	9742
74	TYNGSBOROUGH	237.11	1.12	3969
75	LITTLETON	239.31	1.09	2524
76	NASHOBA	239.68	1.04	5805
77	IPSWICH	237.56	1.01	3733
78	GEORGETOWN	237.24	1.01	2523
79	BROCKTON	222.98	.96	29667
80	MOHAWK TRAIL	232.22	.91	3894
81	GARDNER	229.43	.87	6094
82	AUBURN	236.00	.86	4839
83	NORWELL	241.44	.77	3976
84	HARWICH	234.18	.74	3087
85	GROTON DUNSTABLE	240.58	.70	4674
86	AMESBURY	233.05	.69	5456
87	BERLIN BOYLSTON	239.51	.68	2035
88	GREENFIELD	228.59	.66	4801
89	NEW BEDFORD	222.39	.66	28436
90	GRAFTON	236.90	.65	4384
91	DIGHTON REHOBOTH	235.37	.58	6653
92	LEICESTER	233.67	.52	3861
93	QUABBIN	234.67	.46	6150
94	SPENCER EAST BROOKFIELD	232.37	.37	4516
95	FOXBOROUGH	236.38	.36	5596
96	QUABOAG REGIONAL	230.95	.34	3044
97	SILVER LAKE	235.57	.33	13567
98	CHELMSFORD	238.29	.33	11297
99	HATFIELD	236.96	.32	869
100	KING PHILIP	237.00	.29	9031
101	LYNN	222.50	.28	26397
102	MONSON	234.00	.27	2763

RANK	SCHOOL	98-00	RES890	3 YR TOTAL EXAMS
103	EVERETT	228.95	.23	9509
104	BOURNE	232.03	.23	4912
105	BRIDGEWATER RAYNHAM	234.09	.22	11429
106	FAIRHAVEN	230.09	.15	6249
107	NORTHBRIDGE	231.48	.14	4465
108	HOPEDALE	235.45	.13	2047
109	DUXBURY	240.99	.11	6113
110	QUINCY	231.56	.07	16745
111	HOLBROOK	230.54	.01	2779
112	WAKEFIELD	235.68	.01	6923
113	MIDDLEBOROUGH	230.54	.00	7331
114	WORCESTER	225.62	-.01	42384
115	WINCHESTER	244.00	-.02	6497
116	CANTON	238.13	-.05	5852
117	METHUEN	230.81	-.05	13071
118	BOSTON	220.23	-.06	110732
119	ROCKLAND	231.74	-.06	5353
120	ADAMS CHESHIRE	230.28	-.07	4170
121	EASTON	237.22	-.10	7151
122	HAMPDEN WILBRAHAM	237.14	-.11	7763
123	WAREHAM	228.48	-.17	7319
124	HOPKINTON	239.82	-.21	4795
125	HOLLISTON	237.88	-.34	5925
126	REVERE	226.61	-.38	10765
127	NORTH ANDOVER	237.97	-.38	8377
128	UXBRIDGE	234.14	-.39	4102
129	WALPOLE	237.32	-.39	7249
130	MASHPEE	230.48	-.39	4063
131	FALL RIVER	222.59	-.40	23272
132	RANDOLPH	228.57	-.40	8154
133	NORTH BROOKFIELD	231.03	-.41	1637
134	DUDLEY CHARLTON REGIONAL	233.76	-.43	7752
135	CENTRAL BERKSHIRE	234.09	-.57	5154
136	DANVERS	235.13	-.72	7697
137	NORTH MIDDLESEX	235.74	-.73	9450
138	FITCHBURG	224.87	-.75	10733
139	DARTMOUTH	232.38	-.83	8570
140	OLD ROCHESTER	234.87	-.84	5626
141	NORTH ATTLEBOROUGH	234.99	-.86	8938
142	HAMPSHIRE	234.74	-.91	4693
143	LEE	232.43	-.96	1862
144	LYNNFIELD	239.48	-.99	3940
145	WAYLAND	243.96	-.99	5531
146	MASCONOMET	240.45	-1.00	8017
147	CARVER	232.78	-1.04	4332
148	MALDEN	227.59	-1.05	10110
149	LONGMEADOW	241.74	-1.06	6444
150	BILLERICA	234.17	-1.07	12163
151	COHASSET	241.84	-1.09	2619
152	WHITMAN HANSON	233.31	-1.10	8934
153	SOMERSET	231.81	-1.10	7517
154	ASHBURNHAM WESTMINSTER	234.39	-1.13	5245

RANK	SCHOOL	98-00	RES890	3 YR TOTAL EXAMS
155	GATEWAY	230.39	-1.13	3111
156	BLACKSTONE MILLVILLE	233.46	-1.14	4570
157	MARSHFIELD	236.18	-1.15	8932
158	GLOUCESTER	230.78	-1.16	8475
159	HULL	230.89	-1.16	2886
160	GRANBY	234.13	-1.19	2189
161	NANTUCKET	234.53	-1.20	2468
162	MELROSE	234.40	-1.23	6943
163	ABINGTON	233.40	-1.23	4523
164	HOLYOKE	217.27	-1.23	13769
165	GILL MONTAGUE	229.80	-1.29	3613
166	SUTTON	235.13	-1.31	2966
167	LEOMINSTER	230.11	-1.33	12025
168	BURLINGTON	236.05	-1.35	7009
169	DOUGLAS	232.72	-1.35	2426
170	MENDON UPTON	236.51	-1.35	3709
171	TEWKSBURY	234.24	-1.43	8261
172	HUDSON	233.94	-1.44	4995
173	DEDHAM	234.53	-1.48	5925
174	OXFORD	229.92	-1.51	4275
175	SOMERVILLE	226.10	-1.56	10365
176	SOUTHWICK TOLLAND	232.01	-1.58	4473
177	WEYMOUTH	231.26	-1.61	13580
178	WEST SPRINGFIELD	228.63	-1.62	7948
179	ATHOL ROYALSTON	227.57	-1.64	4614
180	ASHLAND	235.70	-1.71	4349
181	SWANSEA	232.08	-1.81	4719
182	BERKSHIRE HILLS	230.32	-1.84	3803
183	WESTFIELD	228.33	-1.95	13424
184	MEDFORD	228.60	-1.97	9304
185	DRACUT	232.49	-2.00	7946
186	WATERTOWN	232.88	-2.03	5141
187	WINTHROP	231.51	-2.13	4392
188	WEBSTER	228.07	-2.14	3949
189	PITTSFIELD	228.39	-2.17	13760
190	TRITON	233.83	-2.21	6812
191	ATTLEBORO	229.43	-2.22	13550
192	PEABODY	232.03	-2.23	12667
193	MANCHESTER	240.05	-2.23	2057
194	PALMER	228.50	-2.24	4445
195	CHICOPEE	224.26	-2.30	14424
196	EAST BRIDGEWATER	232.35	-2.33	5117
197	LINCOLN SUDBURY	242.64	-2.33	10089
198	PIONEER VALLEY REG.	230.93	-2.37	2389
199	SALEM	227.52	-2.48	9274
200	WILMINGTON	234.59	-2.50	6804
201	WALTHAM	229.51	-2.53	10116
202	WARE	227.63	-2.53	2722
203	MILLIS	234.89	-2.56	2288
204	LUNENBURG	234.15	-2.68	3881
205	SEEKONK	232.46	-2.81	4746
206	LUDLOW	230.30	-2.83	6208
207	FREETOWN LAKEVILLE	231.36	-2.91	6172

RANK	SCHOOL	98-00	RES890	3 YR TOTAL EXAMS
208	MARBLEHEAD	238.98	-2.94	5612
209	BELLINGHAM	232.54	-3.00	5359
210	SOUTH HADLEY	231.13	-3.01	4797
211	SAUGUS	231.55	-3.04	6638
212	WINCHENDON	227.46	-3.09	3995
213	TAUNTON	226.04	-3.33	15658
214	EASTHAMPTON	228.51	-3.49	3731
215	MARLBOROUGH	231.44	-3.55	8028
216	CAMBRIDGE	226.07	-3.67	13788
217	LAWRENCE	217.50	-3.68	21674
218	HAVERHILL	226.92	-4.24	16712
219	MAYNARD	231.77	-4.45	2656
220	AVON	227.79	-4.45	1660
221	LOWELL	222.05	-4.60	29854
222	WESTPORT COMMUNITY	229.16	-5.02	3871
223	NARRAGANSETT	227.87	-5.17	2941
224	SOUTHERN BERKSHIRE	228.34	-6.08	2157
225	DOVER SHERBORN	244.12	-6.11	3760
226	WESTON	245.15	-6.40	4087

Details of the Value Added Model

As explained in the text, this model attempts to estimate the three year grand average score that would be predicted for each district purely on the basis of its economics and demographics. A comparison of that predicted score with the district's actual grand average shows whether the district over- or under- performed its demographics and economics. The residual — the difference between the two scores — quantifies the degree of that over or under-performance.

Formally, the value added model is equivalent to our full model presented in Appendix 4, save that the only covariates this model employs are those for economic or demographic factors. The Ordinary Least Squares version of this model would be:

$$Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \varepsilon_i \qquad \text{(A5.1)}$$

where, Y_i, i = 1,2,...226 is the grand average of MCAS scores for years 1998, 1999, and 2000 for district i, and X_{ij}, j = 1,2,3,4 are the covariates of economic and demographic factors. They are AFRICAN-AMERICAN, PERCAP, TWOPHLD, and TAFDCPER. (LIM.ENG, which might quite reasonably be deemed a non-school related variable, is not used in this equation, since in combination with these variables alone it is not significant.) Not surprisingly, however, a Moran test indicates that the residuals of Equation (A5.1) are also spatially autocorrelated. The test results are summarized in Table A5.1.

Table A5.1

Moran's autocorrelation test for the residuals from an ordinary linear regression model A5.1.

Moran Autocorrelation	Z statistic	P-value
0.2018	5.2931	0.0000

Once again, accordingly, we turn to a Simultaneous Spatial Autoregression (SAR) model equation based on the S PLUS function SLM of Kaluzny and the earlier work of Cliff and Ord.[81] Here the model is:

$$Y_i = \mu_i + \rho\sum_{j=1}^{n} W_{ij}(Y_j - \mu_i) + \varepsilon_i \qquad \text{(A5.2)}$$

where

$$\mu_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} \qquad \text{(A5.3)}$$

The estimated coefficients and p-values are listed in Table A5.2.

Table A5.2

Estimated Coefficients and P-values of Value Added Model

	S-PLUS Estimated Coefficients
INTERCEPT	221.54(.00)
AFRICAN	-0.16(.00)
PERCAP	0.59(.00)
TWOPHLD	0.12(.00)
TAFDCPER	-2.12(.00)

Appendix 6
Improvements in School District Scores 1998–2000
Most to Least
(Average Annual Change – See Text)

SCHOOL	98-00	TOTAL EXAMS	AV. ANNUAL CHANGE
HOLBROOK	230.54	2779	2.80
BRAINTREE	237.78	9794	2.68
SOMERSET	231.81	7517	2.60
CARVER	232.78	4332	2.57
SOUTHWICK TOLLAND	232.01	4473	2.42
SUTTON	235.10	2966	2.23
FRONTIER	238.08	3277	2.21
FOXBOROUGH	236.38	5596	2.15
ABINGTON	233.40	4523	2.13
UXBRIDGE	234.14	4102	2.13
RANDOLPH	228.57	8154	2.13
WESTFORD	242.30	8157	2.12
COHASSET	241.84	2619	2.10
DEDHAM	234.53	5925	2.08
TYNGSBOROUGH	237.11	3969	2.05
NAUSET	238.97	6837	2.03
BOURNE	232.03	4912	2.02
MANSFIELD	236.75	7361	2.00
TAUNTON	226.04	15658	1.95
DOUGLAS	232.72	2426	1.92
MASHPEE	230.48	4063	1.91
CANTON	238.13	5852	1.90
LONGMEADOW	241.74	6444	1.87
WATERTOWN	232.88	5141	1.86
HOPEDALE	235.45	2047	1.85
BEDFORD	241.95	4113	1.84
SOUTHBRIDGE	227.66	4872	1.79
LEICESTER	233.67	3861	1.76
NANTUCKET	234.53	2468	1.72
WARE	227.63	2722	1.72
GRANBY	234.13	2189	1.71
GEORGETOWN	237.24	2523	1.70
SWAMPSCOTT	238.73	5536	1.68
FRANKLIN	239.04	9321	1.67
LYNNFIELD	239.48	3940	1.66
MANCHESTER	240.05	2057	1.66
TANTASQUA	233.47	6797	1.65
PROVINCETOWN	234.00	856	1.62
BLACKSTONE MILLVILLE	233.46	4570	1.60
MONSON	234.00	2763	1.60
EASTON	237.22	7151	1.57
GARDNER	229.43	6094	1.57
NORWELL	241.44	3976	1.55

SCHOOL	98-00	TOTAL EXAMS	AV. ANNUAL CHANGE
SALEM	227.52	9274	1.54
LYNN	222.50	26397	1.53
MILFORD	233.30	7853	1.52
HOPKINTON	239.82	4795	1.48
EAST BRIDGEWATER	232.35	5117	1.48
WAREHAM	228.48	7319	1.48
ROCKPORT	237.06	2343	1.48
MEDFORD	228.60	9304	1.48
NORTHBRIDGE	231.48	4465	1.47
WESTBOROUGH	242.76	6251	1.47
NEEDHAM	243.19	8797	1.45
WAYLAND	243.96	5531	1.44
BURLINGTON	236.05	7009	1.43
ROCKLAND	231.74	5353	1.43
WINTHROP	231.51	4392	1.42
MELROSE	234.40	6943	1.42
IPSWICH	237.56	3733	1.38
CONCORD CARLISLE	245.54	6953	1.38
SHREWSBURY	241.18	8239	1.37
HAMPSHIRE	234.74	4693	1.35
REVERE	226.61	10765	1.32
TEWKSBURY	234.24	8261	1.32
AMESBURY	233.05	5456	1.31
MILTON	237.34	7726	1.28
CHELMSFORD	238.29	11297	1.28
SEEKONK	232.46	4746	1.27
LITTLETON	239.31	2524	1.26
HULL	230.89	2886	1.24
FALMOUTH	233.24	9742	1.21
MILLIS	234.89	2288	1.20
EASTHAMPTON	228.51	3731	1.19
WEST BOYLSTON	238.44	2284	1.19
SOUTHERN BERKSHIRE	228.34	2157	1.18
DANVERS	235.13	7697	1.18
CHATHAM	235.84	1379	1.18
HUDSON	233.94	4995	1.18
MILLBURY	234.44	3542	1.18
BERLIN BOYLSTON	239.51	2035	1.17
STOUGHTON	233.68	8248	1.17
AYER	232.86	3515	1.15
FREETOWN LAKEVILLE	231.36	6172	1.14
BELCHERTOWN	234.58	4618	1.14
BEVERLY	235.51	9693	1.12
BOSTON	220.23	110732	1.11
LEOMINSTER	230.11	12025	1.10
MOHAWK TRAIL	232.22	3894	1.09
ANDOVER	242.29	11851	1.08
HINGHAM	241.70	6934	1.07
BELMONT	243.95	7083	1.06

SCHOOL	98-00	TOTAL EXAMS	AV. ANNUAL CHANGE
OLD ROCHESTER	234.87	5626	1.04
GLOUCESTER	230.78	8475	1.01
GRAFTON	236.90	4384	1.00
NARRAGANSETT	227.87	2941	0.97
NORTON	235.53	5351	0.95
MARBLEHEAD	238.98	5612	0.94
NEW BEDFORD	222.39	28436	0.94
WEST BRIDGEWATER	235.53	2107	0.94
HANOVER	239.09	5112	0.94
WALTHAM	229.51	10116	0.93
EAST LONGMEADOW	237.14	5384	0.93
METHUEN	230.81	13071	0.93
MARTHAS VINEYARD	236.04	4892	0.93
NORTH ANDOVER	237.97	8377	0.93
TRITON	233.83	6812	0.92
CLINTON	232.96	3809	0.92
LOWELL	222.05	29854	0.91
WACHUSETT REG	240.95	13365	0.89
GILL MONTAGUE	229.80	3613	0.88
BARNSTABLE	233.22	13970	0.88
GREENFIELD	228.59	4801	0.87
BROCKTON	222.98	29667	0.87
SILVER LAKE	235.57	13567	0.86
WAKEFIELD	235.68	6923	0.86
DUDLEY CHARLTON REGIONAL	233.76	7752	0.86
FRAMINGHAM	233.61	14584	0.85
QUINCY	231.56	16745	0.84
DIGHTON REHOBOTH	235.37	6653	0.84
ACTON BOXBOROUGH	245.32	10307	0.82
MIDDLEBOROUGH	230.54	7331	0.81
MARLBOROUGH	231.44	8028	0.76
DRACUT	232.49	7946	0.76
HOLLISTON	237.88	5925	0.76
PIONEER VALLEY REG.	230.93	2389	0.75
FAIRHAVEN	230.09	6249	0.74
DUXBURY	240.99	6113	0.74
MARSHFIELD	236.18	8932	0.73
NORTH ADAMS	228.05	5160	0.73
PENTUCKET REGIONAL	239.20	6901	0.73
NORTH ATTLEBOROUGH	234.99	8938	0.73
AGAWAM	232.71	8643	0.72
NASHOBA	239.68	5805	0.72
SWANSEA	232.08	4719	0.71
FALL RIVER	222.59	23272	0.71
WOBURN	237.98	9113	0.70
NORTHBORO SOUTHBORO	242.30	8005	0.69
PEABODY	232.03	12667	0.68
ATTLEBORO	229.43	13550	0.66

SCHOOL	98-00	TOTAL EXAMS	AV. ANNUAL CHANGE
ASHBURNHAM WESTMINSTER	234.39	5245	0.66
HAMPDEN WILBRAHAM	237.14	7763	0.65
WINCHESTER	244.00	6497	0.65
KING PHILIP	237.00	9031	0.63
PLYMOUTH	234.18	17213	0.62
MASCONOMET	240.45	8017	0.61
FITCHBURG	224.87	10733	0.60
MEDWAY	240.05	4785	0.60
WEST SPRINGFIELD	228.63	7948	0.59
NORTHAMPTON	234.23	6101	0.59
WEYMOUTH	231.26	13580	0.59
WESTFIELD	228.33	13424	0.57
BELLINGHAM	232.54	5359	0.57
SCITUATE	238.98	5997	0.57
SPRINGFIELD	219.19	46012	0.56
WILMINGTON	234.59	6804	0.56
WESTWOOD	242.53	4769	0.56
BILLERICA	234.17	12163	0.55
PALMER	228.50	4445	0.53
HAMILTON WENHAM	241.37	5026	0.53
RALPH MAHAR	230.64	3686	0.53
DENNIS YARMOUTH	232.72	8965	0.52
GATEWAY	230.39	3111	0.52
STONEHAM	237.53	5578	0.51
HAVERHILL	226.92	16712	0.48
CHICOPEE	224.26	14424	0.47
NATICK	238.07	7916	0.47
SOMERVILLE	226.10	10365	0.47
NORWOOD	237.57	6848	0.46
READING	240.82	8300	0.46
AVON	227.79	1660	0.45
HARVARD	245.60	2451	0.45
QUABBIN	234.67	6150	0.44
SAUGUS	231.55	6638	0.43
ATHOL ROYALSTON	227.57	4614	0.42
LEXINGTON	244.65	11771	0.40
SOUTH HADLEY	231.13	4797	0.39
BROOKLINE	241.47	11691	0.35
WALPOLE	237.32	7249	0.35
ARLINGTON	239.12	8153	0.34
DARTMOUTH	232.38	8570	0.33
DOVER SHERBORN	244.12	3760	0.32
LUNENBURG	234.15	3881	0.31
SANDWICH	240.02	7946	0.29
NEWBURYPORT	237.17	4954	0.27
NEWTON	243.60	22600	0.26
ASHLAND	235.70	4349	0.26
WHITMAN HANSON	233.31	8934	0.24

SCHOOL	98-00	TOTAL EXAMS	AV. ANNUAL CHANGE
NORTH READING	241.52	4468	0.23
QUABOAG REGIONAL	230.95	3044	0.23
LENOX	239.78	1701	0.23
HADLEY	238.29	1274	0.22
NORTH MIDDLESEX	235.74	9450	0.21
LINCOLN SUDBURY	242.64	10089	0.21
ADAMS CHESHIRE	230.28	4170	0.20
SPENCER EAST BROOKFIELD	232.37	4516	0.18
CHELSEA	222.75	9175	0.17
WESTON	245.15	4087	0.11
MALDEN	227.59	10110	0.09
EVERETT	228.95	9509	0.09
OXFORD	229.92	4275	0.03
WELLESLEY	245.38	6971	0.03
HATFIELD	236.96	869	0.01
MEDFIELD	244.16	5382	-0.06
WINCHENDON	227.46	3995	-0.06
WORCESTER	225.62	42384	-0.06
HOLYOKE	217.27	13769	-0.08
WESTPORT COMMUNITY	229.16	3871	-0.09
LUDLOW	230.30	6208	-0.11
NORTH BROOKFIELD	231.03	1637	-0.11
LAWRENCE	217.50	21674	-0.14
BRIDGEWATER RAYNHAM	234.09	11429	-0.24
GROTON DUNSTABLE	240.58	4674	-0.25
MOUNT GREYLOCK	234.50	3398	-0.32
MENDON UPTON	236.51	3709	-0.40
SHARON	242.45	6936	-0.41
AUBURN	236.00	4839	-0.50
WEBSTER	228.07	3949	-0.50
LEE	232.43	1862	-0.57
PITTSFIELD	228.39	13760	-0.59
MAYNARD	231.77	2656	-0.72
HARWICH	234.18	3087	-0.88
AMHERST PELHAM	237.35	8709	-1.19
CENTRAL BERKSHIRE	234.09	5154	-1.27
BERKSHIRE HILLS	230.32	3803	-2.48
CAMBRIDGE	226.07	13788	-3.14

Appendix 7
Analysis of Change in MCAS Scores, 1998 to 2000

Our main text has emphasized the special problems involving measurement and random error that beset analyses of changes in test scores. Bearing in mind all the necessary qualifications, let Y_{ij} be the grand average of MCAS scores in the i^{th} district, $i = 1,2,\cdots,226$, and j^{th} year, and $T_{ij} = 1,2,3$ for j = 98,99,00. Then the first level model is:

$$Y_{ij} = \beta_{0i} + \beta_{1i} T_{ij} + \varepsilon_{ij} \tag{A7.1}$$

where

$$\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{12}^2 & \sigma_2^2 & \sigma_{23}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_3^2 \end{pmatrix} \right]$$

Here the error term ε_{ij} is assumed to be independent across each district *i* and normally distributed. β_{0i} is the random intercept that varies across districts while β_{1i} is the average rate of change in MCAS scores in district *i* over the three year period.[88] Both the intercept and growth-rate are allowed to vary at level 2 as outcome variables that depend on measured district characteristics. The second level models are:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} X_{i1} + \gamma_{02} X_{i2} + \gamma_{03} X_{i3} + \gamma_{04} X_{i4} + \gamma_{05} X_{i5} + \gamma_{06} X_{i6} + \gamma_{07} X_{i7} + u_{0i} \tag{A7.2}$$

and

$$\beta_{1i} = \gamma_{10} + \gamma_{11} X_{i8} + \gamma_{12} X_{i9} \tag{A7.3}$$

where X_{ij} are the district characteristic variables associated with the coefficients listed in Table A7.1 and $u_{0i} \sim N(0, \tau)$ Substituting Equations (A7.2) and (A7.3) in Equation (A7.1) leads to the full model:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01} X_{i1} + \gamma_{02} X_{i2} + \gamma_{03} X_{i3} + \gamma_{04} X_{i4} + \gamma_{05} X_{i5} + \gamma_{06} X_{i6} + \gamma_{07} X_{i7} \\ & + \gamma_{10} T_{ij} + \gamma_{11} T_{ij} X_{i8} + \gamma_{12} T_{ij} X_{i9} + u_{0i} + \varepsilon_{ij} \end{aligned} \tag{A7.4}$$

The estimated coefficients are listed in Table A7.1.

We estimated this model using both the HMLM component of the HLM program and SAS PROC MIXED. In HLM, we estimated an unrestricted model, since comparisons with various alternatives indicated that it performed better. The deviance statistics show clearly that a substantial amount of the improvement in our model comes from the intercept, though the T statistics for the terms for both state aid and district changes in median household income between 1989 and 1999 are fine.

Table A7.1

District characteristics variables with estimated coefficients and P-values.

Fixed Effect	Definition	Coefficient (P-value)	S.E.
γ_{00} = INTERCEPT		212.14 (.00)	2.17
γ_{01} = TPERCAP	Per Capita Income (Unit: $1,000)	.49 (.00)	.05
γ_{02} = TWOPHLD	Percent of Households With 2 Parents	.06 (.02)	.03
γ_{03} = TAFDCPER	Percentage of TAFDC	-1.82 (.00)	.29
γ_{04} = AFRICAN-AMERICAN	Percentage of Students of African-American Descent	-0.22 (.00)	.03
γ_{05} = TSMX989	Teachers Maximum Salary 1998-99 (Unit: $1,000)	.15 (.00)	.05
γ_{06} = TSAL	Superintendent's Salary (Unit: $1000)	.04 (.00)	.01
γ_{07} = LIM.ENG	Percentage of Students with Limited English	-0.16 (.00)	.06
γ_{10} = time	YEAR	.46 (.00)	.08
γ_{11} = GRC7TS40	Difference in Chap 70 Funds As Percent of Actual Total Spending 1994 - 2000	1.86 (.01)	.69
γ_{12} = *PCHHIN*98	Percent Change in Median Household Income 1989-1999	1.41 (.01)	.50